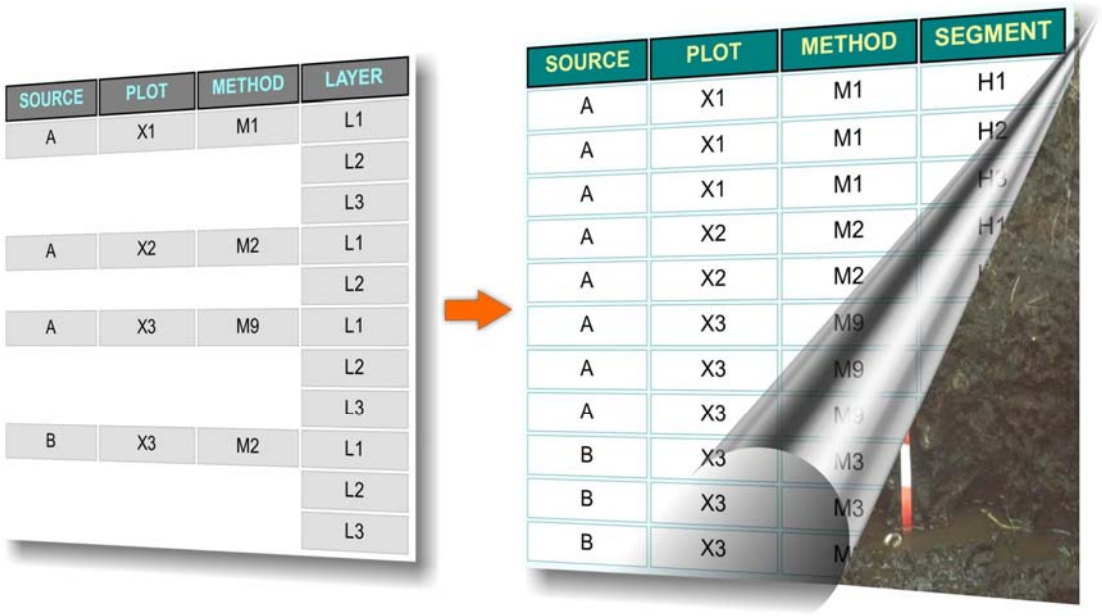
JRC Scientific and Technical Reports

Data Update and Model Revision for Soil Profile Analytical Database of Europe of Measured Parameters (SPADE/M2)

Roland Hiederer



EUR 24333EN - 2010

JRC EUROPEAN COMMISSION

Institute for Environment and Sustainability

The mission of the JRC-IES is to provide scientific-technical support to the European Union's policies for the protection and sustainable development of the European and global environment.

European Commission
Joint Research Centre
Institute for Environment and Sustainability

Contact information

R. Hiederer
European Commission
Joint Research Centre
Institute for Environment and Sustainability
Via Enrico Fermi, 2749 - 21027 - Ispra (VA) – Italy
E-mail: roland.hiederer@jrc.ec.europa.eu

http://ies.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

Legal Notice

JRC Catalogue number: LB-NA-24333-EN-C

EUR 24333 EN
ISBN 978-92-79-15646-5
ISSN 1018-5593
DOI 10.2788/85262

Luxembourg: Office for Official Publications of the European Communities

Printed in Italy

This document may be cited as follows:

Hiederer, R. (2010) Data Update and Model Revision for Soil Profile Analytical Database of Europe of Measured Parameters (SPADE/M2). EUR 24333 EN. Luxembourg: Office for Official Publications of the European Communities. 55pp.

European Commission Joint Research Centre
Institute for Environment and Sustainability
TP 261
21027 Ispra (VA)
Italy

Table of Contents

Page

List of Figures

List of Tables

Page

List of Acronyms

ACRONYM	TEXT
ANSI	American National Standards Institute
BLOB	Binary large objects
DBMS	Database management system
FAO	Food and Agriculture Organization of the United Nations
ICP Forest	International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests
ID	Identifier
IGN	Institut Géographique National
NSRI	National Soil Resources Institute, Cranfield University
NTF	Nouvelle Triangulation de la France
OC	Organic carbon
OLE	Object linking and embedding
OM	Organic matter
OS	Operating system
PTR	Pedo-Transfer Rule
RAM	Random Access Memory
RDBMS	Relational database management system
SGDBE	Soil Geographic Database of Eurasia
SPADE	Soil Profile Analytical Database of Europe
SPADE/M	Soil Profile Analytical Database of Europe of Measured parameters
SQL	Structured Query Language
WGS84	World Geodetic System datum 84

1 INTRODUCTION

The *Soil Profile Analytical Database of Europe of Measured* parameters (SPADE/M) is part of the distribution package of the *Soil Geographic Database of Eurasia* (SGDBE). It was created to provide a common structure for storing standardized information on typical soil profile properties of major European soils. The criterion applied to include a profile in the database was to adequately cover the range of European soils rather than to sample profiles for a statistical representation of soil properties at European scale. The typical combinations of profile parameters and morphological characteristics of the sample site were intended to support the definition of generalized rules for estimating pedological and hydrological properties of the *pedo-transfer rule* (PTR) database of the SGDBE. Compared to the spatial database the information on measured profiles has received relatively little attention. Since the incorporation of the profile database into the SGDBE and the release of the database in 1999 it has undergone only one change when in 2005 the information stored in separate files was transferred to a standardized format and stored in a database (Hiederer, *et al*., 2006). In 2008 original hard-copies on profile measurements were re-discovered at the *National Soil Resources Institute, Cranfield University* (NSRI). The values recorded thereon were found to differ from data of the SGDBE. To make the original data more generally available the profiles were added to the existing database. This step required changes to the structure of the database and a validation of the all entries for accurate and reliable data storage and retrieval.

1.1 Recovered Measured Profile Data

For soil profile data to be useful for the verification of spatial data the profiles the geographic position of the profiles is generally needed. In SAPDE/M coordinates on the position of profiles could be included for 408 out of 496 plots. For 385 plots the positions could be transformed to a geographic coordinate system. For the UK more than 50% of the soil profiles were reported without information on plot positions. This lack of information on the position of the plots very much reduces the value of the profile data for verifying spatial soil data. With the creation of SPADE/M attempts were undertaken to recover the missing coordinates of UK plots. In 2008 data on 64 Proforma II soil profiles, as prepared by NSRI and preceding the processed profiles included in the SGDBE database, could be recovered in form of hard-copies. All recovered former profile data were for plots located in England and Wales. The additional information could be linked to existing profiles by the code used for the location name, which was handwritten on all hard-copies. The printouts include data for 52 profiles for which data were already reported in SPADE/M and 12 profiles for which no correspondence to an existing profile could be found.

The data of the recovered plot and profiles have been transferred to an electronic format by manually entering the values into a database. The new entries were checked together with the previous data of SPADE/M using the verification method outline later in this document. All profile data were subsequently stored in the restructured SPADE/M2 database.

1.2 Changes to SPADE/M Data Model

To make wider use of the recovered measured profile data the values recorded on the hardcopies were entered into the SPADE/M tables. It was intended not to replace the previously stored data but to keep the information of the SGDBE for reference and to add the recovered information. This approach required changes to the data model because the original data model did not foresee storing data of more than one profile for a sampling site. Since changes to the database were inevitable it was decided to completely restructure the SPADE/M data model to align it with more recent models for storing soil profile data. The new data model arranges data more similar to the structures used by the re-engineered data model of the Soil Condition survey performed on Forest Focus/ICP Forests Level 1 sites and the structure used to store soil data from the BioSoil demonstration project. The data model moves away from the arrangement of explicitly declaring a field for each parameter, which is familiar and resembles the tabular arrangement of data stored in a spreadsheet. While the new structure is more efficient for storing soil profile data, more flexible for integrating data from different surveys, and allows better control of data formats and consistency of values, it is not well suited to be used by spreadsheet packages without pre-processing.

1.3 Data Verification

With new data entering the database aspects of verifying the data values of the recovered, but also the existing, profiles became relevant. Methods of verifying soil profile data have been used in other projects, such as the Forest Focus monitoring activity (Hiederer, *et al*., 2007) and the BioSoil demonstration project (Durrant-Houston & Hiederer, 2009).

The aim of the data verification is to define a transparent approach to providing accurate and reliable data for use as a reference in modeling activities where soil properties are estimated, such as generating spatial layers. For the measured profile data the verification concerns checks on data coherence and acceptable ranges for the values reported for a parameter. Limited checks were performed on cross-checking parameters. Not part of the checks was the assessment of the consistency of soil classification schemes with the values reported.

2 MEASURED PROFILE DATA

The details on compiling the data of measured soil profiles as part of the SGDBE are presented elsewhere (Hiederer, *et al*., 2006; Breuning-Madsen & Jones, 1998). The data of measured profiles were assembled by national experts in form of paper notes or computer files. The national data were then transferred to spreadsheet files of a common layout. These files were then distributed between the project partners for further processing and use. The stages of the process are graphically presented in Figure 1.

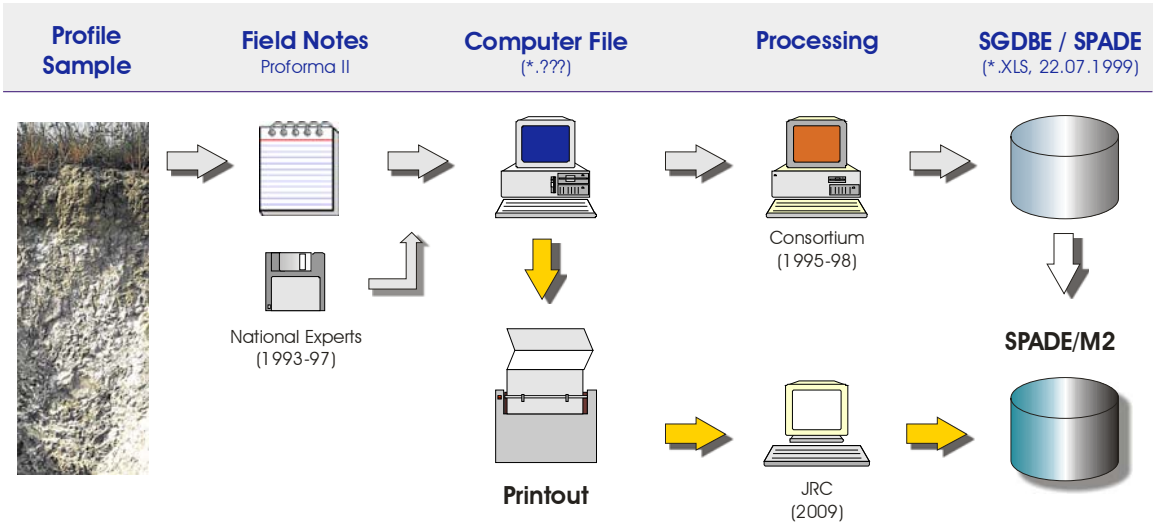


Figure 1: Major Processing Steps of Measured Soil Profile Data

A copy of the files being returned at the end of the project was included in previous versions of the SGDBE. In the course of the search for information on the geographic position of the sites located in the UK paper copies of the data stored in the original spreadsheet files were found at NSRI (J. Hollis & R. Jones, personal communication). These copies could be saved from being disposed of and were used to verify the data stored in the spreadsheets. Significant differences between the data stored on the electronic files and the hardcopies became apparent for many profiles. From subsequent inquiries it appears that the data stored on computer files at the end of the project and included in the SGDBE did not in all cases contain the original data provided by the national experts. For several profiles in the UK the data returned seem to be the typified data as intended to be used in the Proforma I format (estimated data; Breuning-Madsen & Jones, 1995) for SPADE.

The exact steps taken to process the original data and any amendments applied to them could not be established after so many years. It has to be assumed that the recovered hardcopies contain the data of the original spreadsheet files, which have to be assumed lost.

2.1 Locations of Recovered Profiles

In the SDGBE the database on measured profiles contains data on 86 plots in the UK. Coordinates are available for 37 plots. Of these, 26 are positioned in Scotland, 11 in England and none in Wales and Northern Ireland. The location of the plots of the former profiles in relation to those of the SGDBE and SPADE/M database are shown in Figure 2.

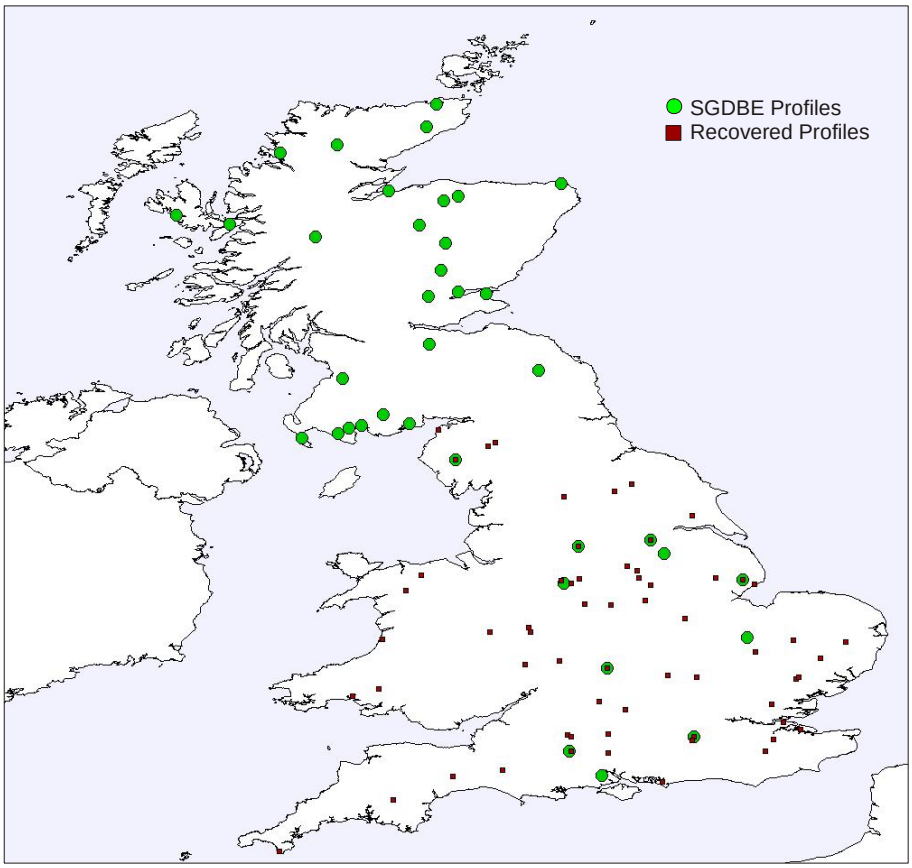


Figure 2: Location of Recovered Profiles and Plots of Profile Database

A visual examination of the graph shows that close proximity between plot positions exists for 8 out of the 11 plots in England. Despite the name the data for England and Wales in the SPADE/M database describe representative profiles whereas the recovered profile data are taken from actual measurements. It is thus not unexpected to find only limited correspondence of plot coordinates between the two data sets. The correspondence between the recovered and the SGDBE data for plot parameters was further evaluated.

2.2 Correspondence of Plot Parameters

Data of the parameters characterizing the plot from the recovered profiles were compared to the corresponding values of the SGDBE. A summary of the comparisons made is given in Table 1.

Table 1: Difference between SGDBE and Recovered Plot Parameter Data

Plot ID	Case ID	Parameter	SGDBE	Recovered	Comment
447	507	COOR_X	-1.9553	-1.9553	
452	513	COOR_X	-1.8922	-1.8756	
489	552	COOR_X	-1.5239	-1.0975	
495	557	COOR_X	0.1517	0.2508	
447	507	COOR_Y	53.0750	54.0750	Possible typing error
452	513	COOR_Y	51.1469	51.1469	
489	552	COOR_Y	50.8608	53.1442	
495	557	COOR_Y	52.4564	52.2933	
482	544	SOIL	lc	Lc-3ac	Nordrach series recorded
487	549	SOIL	lgs	Lgs-2/4a	Ragdale series recorded
All	All	ORIG	1	2	

For most parameters there are only minor spelling differences of location or soil names, with the recovered data being generally more elaborate. In spite of this, there are also instances where the recovered data show considerable variations for the horizon data:

- ***Plot Coordinates***

  For 4 plots differences between the data were found of up to 2.3 deg. In most cases where differences were found they are not minor. In one case a typing error could be responsible (Plot code: 1033, values checked on hardcopy and SGDBE file), but this is very unlikely in other cases. The proportion of inconsistencies in plot coordinates (4/11) casts doubts over the reliability of the plot coordinates. Without a reliable reference the correctness of any of the data cannot be established.

- ***Soil Name***

  Soil names are almost identical between the data sets when the series recorded in the recovered data is not considered. In the SGDBE the soil series names are provided in a separate table (UKCOORD.XLS). The differences found between

the hardcopies and the files were attributed to using lower-case letters in the SGDBE data.

- ***Origin***

  The main difference between the two data sets is expressed in the "*Origin*" of the profile data. The data of the SGDBE describe single representative profiles for England and Wales. In contrast, the recovered profile data originate from an average of a number of profiles. In two cases an origin was given in the recovered data where no value was given in the SGDBE data. It can be safely assumed that also in those cases the data in the SGDBE profiles characterize representative profiles

- ***Altitude, Ground Water Level, Land Use, Parent Material, Depth Ranges***

  For these parameters data were only available from the recovered data of England and Wales.

The two sets of data show almost identical values for profile soil names, but some non-negligible differences for plot coordinates. Compared to the recovered data the parameters stored in the SGDBE are less complete in providing data of the plot parameters.

2.3 Correspondence of Horizon Parameters

Comparing data for horizon parameters between the two data sets is more complex than comparing data for plot parameters because the horizon depth limits do not correspond between the two data sets for the same plot. In an attempt to provide some measure of comparability a standard depth interval was defined for a topsoil layer extending from 0 – 30 cm. Parameter values were interpolated for the layer by weighting horizon values by the proportion of the horizon within the layer. Only those profiles were used in the comparison for which sufficient data were available to completely cover the topsoil layer. Interpolated means for the topsoil layer were computed for key parameters of the horizon.

- ***Texture***

  Of the 6 texture parameters clay content was used to investigate the correlation between the two data sets. The interpolated values for the generally available clay content (46 profiles included) are plotted in Figure 3.

  The graph shows a close correlation between the data ($r^2 = 0.92$), but also some notable differences for profiles with clay content > 60%. For three plots (PLOT_ID: 437, 443, 469) the topsoil profile contains 20% more clay in the recovered data than is found in the processed data of the SGDBE. The profiles

have different soil types (*Jeg*, *Ges*, *Bgc*) and are similar in depth intervals. No information on the origins of the differences for those 3 plots could be identified.

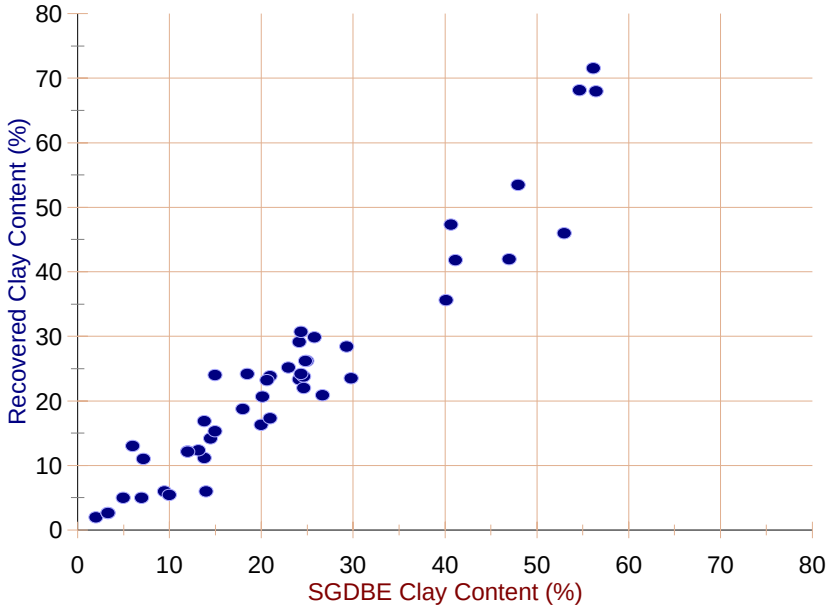


Figure 3: Interpolated Topsoil (0-30cm) Clay Content for SGDBE and Recovered Profiles

- ***Organic Carbon OC), Organic Matter (OM)***

  In the recovered data organic carbon is given instead of organic matter. OC is the measured parameter, while OM is derived from it by applying a conversion factor. That OC instead of OM was measured is indicated the measurement method (Walkley & Black) assigned to the OM values. The conversion factor is widely based on the assumption that organic matter contains 58% organic carbon (Howard & Howard, 1990). Comparing OM from the SGDBE profile data with OC from the recovered profiles should result in a correlation with a coefficient of approx. 1.72. The estimates OM and OC contents for the topsoil layer are graphically presented in Figure 4.

  Interpolated OM and OC values for the topsoil layer could be related for 29 profiles. The regression coefficient is defined by just one profile (PLOT_ID: 495) of a *histosol* (*Oe*) which has a very high degree of leverage on the regression coefficient. In the data the OM content of the first horizon (0-25 cm) is given as 33.0%. The OC content for the first horizon (0-23 cm) is given as 30.6%. Excluding the point from computing predictor values leads to no correlation between OC and OM data. The lack of correlation between OC and

OM can be considered atypical, given that close correlations have been found for other horizon parameters.

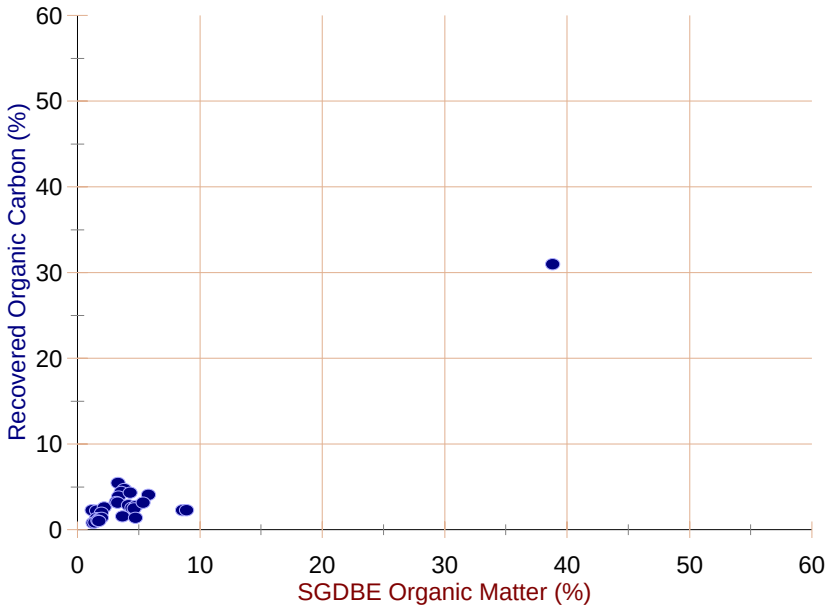


Figure 4: Interpolated Topsoil (0-30cm) OM Content for SGDBE and OC Content for Recovered Profiles

- ***Soil Water Retention (WK)***

  Estimates of the soil water retention for the topsoil layer were computed for 1st soil water retention value (WK1). A graph of the relationship between the SGDBE and the recovered data is presented in Figure 5.

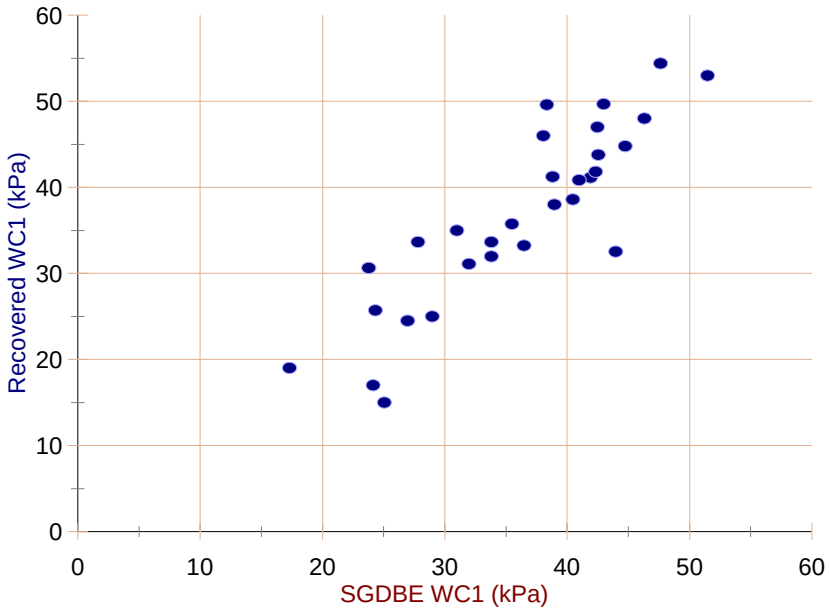


Figure 5: Interpolated Topsoil (0-30cm) Soil Water Retention for SGDBE and Recovered Profiles

A relationship could be established for 30 profiles with a regression coefficient of 1.02 (STD_ERR: 0.025) and a coefficient of determination of (0.77). Thus, the data are highly correlated with a regression coefficient which includes 1.0 within 1 standard error of the coefficient.

- ***Bulk Density (BD)***

  Measured values of bulk density in soil profiles are needed to assess stocks of OC in the soil. A comparison of the interpolated topsoil bulk density between SGDBE and recovered data is presented in Figure 6.

  Interpolated bulk density from 21 profiles could be correlated. The number is comparatively low since for the majority of profiles bulk density is not or only partially available. A linear regression indicates a high correlation (coeff. determ.: 0.72) and a coefficient close to 1.0 (0.97; STD_ERR: 0.03). The relationship is largely determined by two profiles (PLOT_ID: 465, 495). The latter plot describes a profile of an organic soil, for which a discrepancy between the two data sets was already noted when assessing OC vs. OM contents.

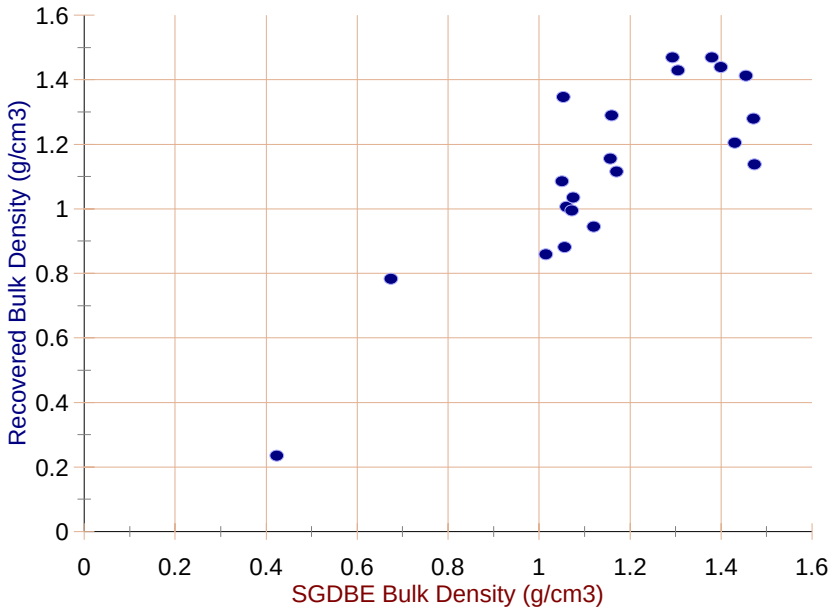


Figure 6: Interpolated Topsoil (0-30cm) Bulk Density for SGDBE and Recovered Profiles

- ***Horizon Name, Gravel, Structure, CaCO3, pH, EXCH_MG, EXCH_K, EXCH_NA, CEC, BS,***

  Data on these horizon parameters were only recoded in the recovered data. The profiles of the SGDBE contained empty entries or an entry of -1.

- ***Nitrate, $CaSO_4$, Electric Conductivity, SAR, ESP, EXCH_CA,***

  For there parameters no or very few values are recorded in either data set.

While there is evidence, other than the code used in the field on data origin, which supports treating the recovered data as representing actual measurements it is not obvious how the data included in the SGDBE were generated. Linear weighting of a parameter by depth range was found to allow a measure of comparison between the recovered data of a single representative profile and the average of a number of profiles of the SGDBE.

2.4 Modifications to French Plot Coordinates

In the preparation of the SPADE/M database the coordinates of 21 plots in France could not be determined with confidence, although coordinates were provided. The original tables included as information the name of the Lambert Conformal Conic projections used in France (Lambert I, Lambert II and Lambert III; Lambert IV for Corsica does not appear). When mapping the plots according to the parameters of the *Nouvelle Triangulation de la France* (NTF) several plots are completely outside France. From the clustering of the plots and the coordinate values it would appear that some plot coordinates were recorded according to the NTF(1922) standard system while other plot coordinates correspond to the zones "Carto" in use by the *Institut Géographique National* (IGN) since 1972. In this system the y-coordinates are preceded by the zone, effectively adding 1,000, 2,000 or 3,000 km to the value (Bouron, 2005).

The coordinates of the sites outside the specified zones were adjusted according to the specifications of the cartographic zone used and the standard projection parameters. They were subsequently converted to geographic coordinates (datum: WGS84). The positions of the 21 French plots with recomputed geographic coordinates for the plots with "Carto" reference are displayed in Figure 7.

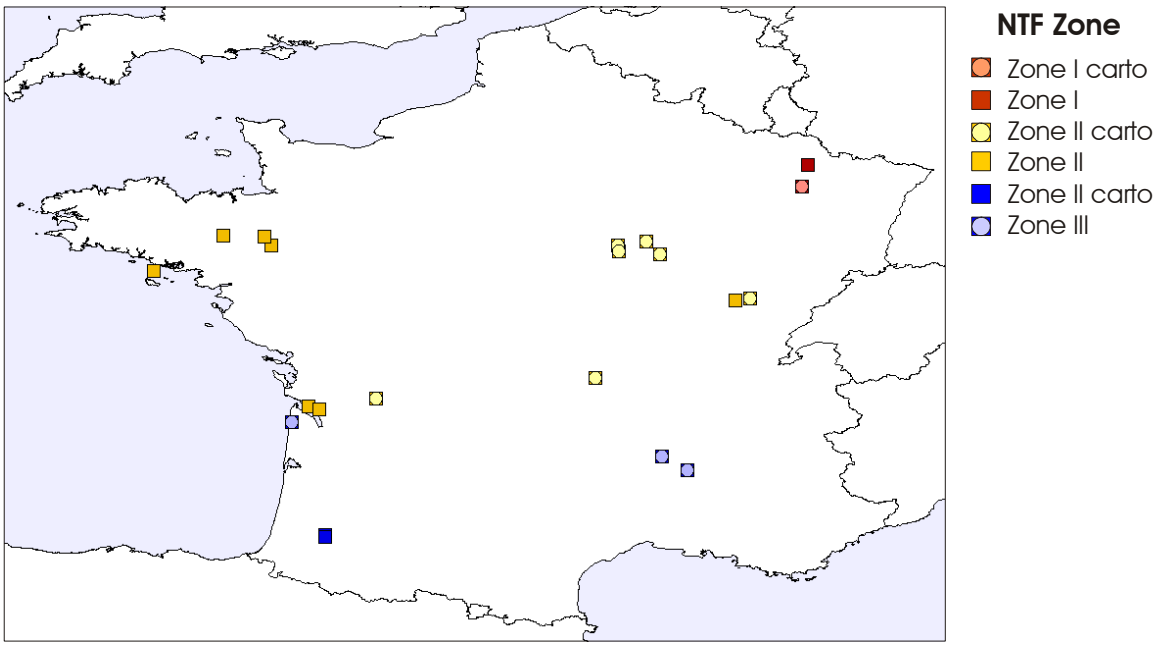


Figure 7: Position of French Plots added to SPADE/M2 Database

The positioning of the plots corresponds to the information of the SGDBE spatial information, although the identifier for the soil mapping unit (SMU) given in the original files could not be reliably used, probably because it refers to an early version of the database. The plot off the coast of Brittany is position on the Quiberon peninsula,

not in the sea. The coordinate adjustment applied to the French plots increased the number of plots with geographic coordinates in the database to 466.

3 DATABASE MODEL

The terminology used to describe database elements and models tends to be obscure and depends to some degree on the database management system (DBMS) used. At the level of planning a database terms such as entity and attributes are used which are translated into tables and fields when implementing the database. The same word may also carry subtle connotations and differences for different DBMS, e.g. field types, domains and concepts of data integrity. For the SPADE/M2 database model only the relational model is considered and a generic and at times simplistic vocabulary was used to describe the database model where possible.

3.1 Basic Database Concepts Applied

In a relational database data are stored in tables (entities). The table columns are fields (attributes), while the rows are records (entity instance). At the intersection of a field and a record data are stored. The arrangement resembles the layout of a spreadsheet, where the storage locations are cells. However, for each field of a database table a *type* has to be declared. All data stored under the field are stored according to the type, whereas in a spreadsheet field types can be declared by cell. General field types are described in Table 2.

Table 2: General Field Types

Field Type	Description
Character	Containing strings of characters and/or numbers.
Numeric	Number field holding values with or without decimals.
Logical	Holds either of two states, True/ False or 1/0.
Date*	Date as date value instead of text string.
BLOB	Binary large object.

* Time fields are not considered separately.

Although there are common types for fields, mainly character, numeric and date, the details of field type specifications differ between DBMS and type dimensions depend to some degree also on the operating system (OS) used when a DBMS is available for several platforms.

- ***Character***

  The field type containing strings is at times referred to as alpha-numeric, string or character (ANSI SQL: *CHARACTER*). Some DBMS distinguish between fixed and variable length character fields. The maximum length of the character field varies and where compatibility is an issue the field dimension should not exceed 254 characters (maximum size in dBASE format[1]). Other DBMS use 255 or more as maximum length.

  For text originating from multi-national sources the country-specific language sets may have to be considered, e.g. for plot names or comments. Apart from the available characters the language driver can also influence the sort order.

- ***Numeric***

  There is a wide range of formats for fields containing numbers and a further distinction into integer and float types should be made:

  Numeric Integer

  Fields of type integer hold whole numbers. Some DBMS distinguish between different types of integer values and the corresponding types can hold values from 2^8 (0-255, 1 byte) to $\pm 2^{31}$ (ANSI SQL: *INTEGER*; -2,147,483,648 to 2,147,483,647, 4 byte). An integer field using 2 bytes (ANSI SQL: *SMALLINT*) can hold values between – 32,768 and 32,767. For dBASE format files a *float* is used to hold integer data.

  Numeric Float

  Numeric values with fractional parts are stored in type *float* or *real*. The field type can be of fixed or floating format, depending on the definition of the position of the decimal point. The range of numbers the field can store and the precision varies greatly: generally available is a single-precision float type (ANSI SQL: *REAL*), which occupies 32 bits and is accurate to 7 decimal digits. The dBASE binary floating-point format can contain up to 20 digits, but due to rounding during calculations the field should be used to store values only up to two decimal places, although the field type can store data with higher precision. For more precision the binary-coded decimal format should be used. Double-precision float formats (ANSI SQL: *FLOAT*) and can store very large values at with higher precision at the price of storage space (8 bytes). Some DBMS allow specifying the number of decimals for numeric data and care should be taken when dimensioning the field for including the sign and decimal separator in the count.

[1] dBASE was originally published by Ashton-Tate, Torrance, California, USA (now defunct)
dBASE plus is a trademark of dataBased Intelligence, Inc., 2548 Vestal Parkway East, Vestal, NY 13850

- ***Logical***

  A logical or bit type field can be seen as the most basic storage type. It can hold data with a binary status, expressed as either *T*(rue) / *F*(alse), *Y*(es) / *N*(o) or *1*/*0*. The conversion of the bit status can be confusing when a system displays one set of the binary status but requires entering a 0 or 1 when modifying the field or, to further confuse matters, a value of -1 to indicate the status *True*. A particularity of some DBMS is that no empty entries are permitted for the field or such a restriction is imposed when importing data. Thus, when importing data into a field of type logical all empty entries are automatically set to *False*. This behaviour of the system can unintentionally bias imported data with null entries.

- ***Date***

  Date fields (ANSI SQL: *DATE*) can display dates in a variety of formats, but in dBASE are always 8 positions wide. Other DBMS use different conventions, such as a combined date and time format. Some DBMS provide date field types which use as format the settings as defined for the OS. This can lead to confusion when exporting the date and importing the data in an environment with different settings.

- ***BLOB***

  Binary large objects can contain a variety of data, such as binary, memo, graphic, or OLE data, which can be sound, documents, images or any other data in binary format. Depending on the DBMS they may be stored as part of a table or separately. The BLOB data type is not considered relevant for this survey.

There is a wide range of variations of numeric fields and additional field types, e.g. auto-increment or sequential, money fields, memo, image, etc. Not all formats are generally available or can be converted between systems. When distributing data the limits in the range of values, in particular for integer and character fields, should be set conservatively. It could be prudent to select the measurement unit of a measurement to allow storing the data within ± 7 digits around the decimal point. For soil data the precision offered should be largely sufficient.

The range of possible values a field can contain is referred to as the *domain*. In early days of relational database the domain use to describe what is now referred to as the field type (Darwen, 2009). The domain definition can be very detailed and conceptually resembles a user-defined type. The meaning of the *data domain* is now closely linked with database integrity. The domain information is largely, but not comprehensively, stored in the tables of the SPADE/M database. Field type, dimensions, formats and ranges are stored, although constraints defined in procedures, such as default values or attribute dependencies, are not included.

Empty fields should receive special mentioning. When a field is empty it contains a *Null* value. Depending on the settings of the relational DBMS (RDBMS) in calculations the

value *NULL* can be either ignored or treated as zero (0). It is therefore advisable to avoid the presence of empty fields in the database where fields are calculated.

Absence of a Null value in a field is also a prerequisite for using the field as a *primary key*. A primary key uniquely identifies a record and consists of a single field or a combination of fields. Keys are central to relational databases because they allow setting relations between tables. Relations are set on fields common to the tables. To identify a value unambiguously one of the common fields must be a primary key of a table. The requirement of precision of the values stored in a key is one of the reasons why a field of type float cannot be used to form primary keys.

3.2 Data Normalization

For relational database designs aspects of normalization are of major concern. The objective of normalization is to allow general-purpose querying and to avoid loss of data integrity from modifying the data (insert, update and delete) (Codd, 1970; Wikipedia, 2010[2]). A normalized database is achieved when in 3rd normal form. Higher forms of normalization are distinguished but were considered not relevant to the re-design of SPADE/M.

An example of arranging data from measured soil profiles of the Proforma II Soil Analytical Data in a table is given in Table 3.

Table 3: Generic Format for Proforma II Soil Analytical Data

Plot Name	Country Code	Country Name	Soil Name	Horizon 1 Name	Horizon 1 Depth	Horizon 1 Clay
Plot A	AX	Xland	Bef-3	Ahg	0-15	62
My Plot	BZ	Ystates	Wallasea	Ah	0-10	54
A Plot	BZ	Ystates	Fladbury	Ahg	0-25	19

Horizon 2 Name	Horizon 2 Depth	Horizon 2 Clay	Horizon 3 Name	Horizon 3 Depth	Horizon 3 Clay
Bg1	15-33	74	Bg2	33-62	61
Bg1	10-35	55			
Eb(g)	25-42	20	Btg	42-53	26

[2] http://en.wikipedia.org/wiki/Data_normalization. Accessed 16.01.2010

The process of normalizing the SPADE data largely follows the steps outlined in the introductory article by Gilfillan (2002).

3.2.1 First Normal Form (1NF)

The aim of the 1st normal form is to eliminate duplicate data and reduce redundancy. For a database to be of 1st normal form it should:

- contain *no repeating groups or sets* (attributes are single values);
- define all *key attributes*;
- have all *attributes dependent on the primary key*.

For each sample plot several attributes are recorded describing the plot. For each plot a single profile is described which consists of several horizons. For each horizon a series of parameters describing the soil conditions are recorded. In the arrangement shown in the generic table (Table 3) reporting horizon and parameter series as an individual field leads to *repeating groups or sets* of the horizon data. Some fields of the repeating group may be empty when fewer horizons were identified in the profile than fields are defined in the table.

To avoid *repeating sets* of data the repeating horizon elements are declared as fields and the horizon name is used as the key attribute[3]. Data on the plot are filled in to complete the records in order for a field to be used as part of a table key. Because the plot name may not be unique a field with a unique identifier for the plot is added to be used as a key field. For the primary key a combination of fields is needed to uniquely identify a record. The resulting table in 1NF is given in Table 4.

[3] This arrangement was already used in the Proforma II spreadsheet data and the situation shown here only serves as an example for repeating groups.

Table 4: Proforma II Soil Analytical Data in 1NF

Plot ID	*Plot Name*	Country Code	Country Name	Soil Name	*Horizon Name*	Depth Range	Clay Content
001	Plot A	AX	Xland	Bef-3	***Ahg***	0-15	62
001	Plot A	AX	Xland	Bef-3	***Bg1***	15-33	74
001	Plot A	AX	Xland	Bef-3	***Bg2***	33-62	61
002	My Plot	BZ	Ystates	Wallasea	***Ah***	0-10	54
002	My Plot	BZ	Ystates	Wallasea	***Bg1***	10-26	55
003	A Plot	BZ	Ystates	Fladbury	***Ahg***	0-25	19
003	A Plot	BZ	Ystates	Fladbury	***Eb(g)***	25-42	20
003	A Plot	BZ	Ystates	Fladbury	***Btg***	42-53	26

All records are now filled with non-null values in the plot and horizon fields. The key fields are "*Plot Name*" and "*Horizon Name*". From the combination of these the primary key can be defined. It is assumed that the data in the field "*Horizon Name*" only appears once in a plot profile.

3.2.2 Second Normal Form (2NF)

The 2nd normal form aims at removing all fields that are not defined by the primary key. For database to be in 2NF complying with the following conditions apply:

- the table is in 1NF;
- the table does not contain *partial dependencies*.

A partial dependency exists when a field depends on only part of the primary key instead of on the entire primary key. In the Proforma II data any data related to the plot depends solely on the "*Plot Name*". To remove partial dependencies the table is split into two tables, one containing the plot attributes and one containing the attributes related to the profile horizons. The new plot table stores a single value for each field and thus eliminates duplication of plot data in the table.

The second component of the primary key ("*Horizon Name*") does not contain partial dependencies. To relate the table to the data on the plot the "*Plot ID*" field is retained. Combined with the field "*Horizon Name*" it forms the primary key of the table.

The structure of the two tables is given in Table 5.

Table 5: Proforma II Soil Analytical Data in 2NF

Plot ID	Plot Name	Country Code	Country Name	Soil Name
001	Plot A	AX	Xland	Bef-3
002	My Plot	BZ	Ystates	Wallasea
003	A Plot	BZ	Ystates	Fladbury

PLOT Table

Plot ID	*Horizon Name*	Depth Range	Clay Content
001	***Ahg***	0-15	62
001	***Bg1***	15-33	74
001	***Bg2***	33-62	61
002	***Ah***	0-10	54
002	***Bg1***	10-26	55
003	***Ahg***	0-25	19
003	***Eb(g)***	25-42	20
003	***Btg***	42-53	26

HORIZON Table

The primary key for the PLOT table is a single field ("*Plot Name*"), while the primary key for the HORIZON table is a combination of two fields, one of which is the primary key of the PLOT table ("*Plot Name*" & "*Horizon Name*").

3.2.3 Third Normal Form (3NF)

The aim of the 3rd normal form is remove fields which are defined by fields other than the primary key. The conditions defining the 3rd normal form are:

- the table is in 2NF;
- the table does not contain *transitive dependencies*.

A transitive dependency exists when a field not part of a key depends on another non-key field rather than directly on the primary key. As the tables are defined in the example entries in the field "*Country Name*" depend on the field "*Country Code*".

The information on the country is duplicated in the table, which leads to larger storage requirements, but also to potentially incongruous situations, for example when a country name is spelled differently in the field. The data is therefore moved to a separate table in which the country name is uniquely defined for each country code. This arrangement allows using the "*COUNTRY Table*" as a look-up table for the values of the field "*Country Code*" in the PLOT Table when entering data so only valid codes are entered and correctly related to the country name. For a more controlling relationship referential

integrity between the PLOT and the COUNTRY tables can be specified on the field in the database. The logical model of the database is presented in Figure 8.

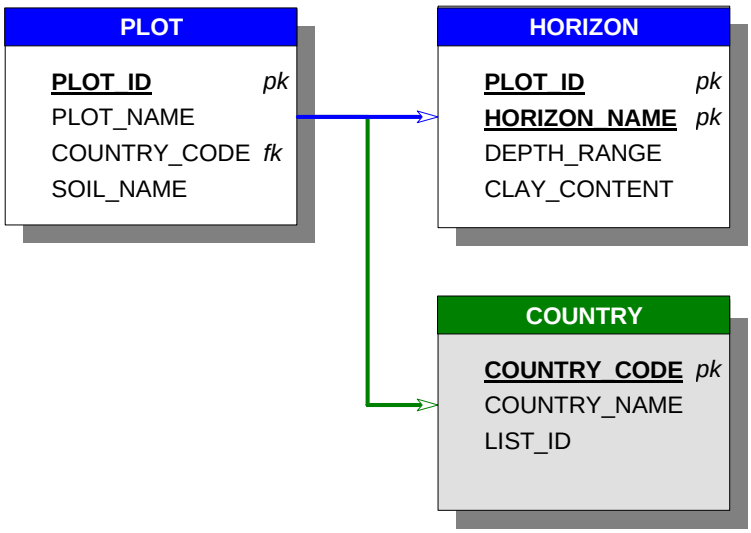


Figure 8: Conceptual Model of Example Proforma II Soil Analytical Data in 3NF

The graph of the conceptual model shows the primary keys of the tables on which the relationships are founded. The "*Country Code*" in the PLOT table forms the foreign key in the relation to the data of the COUNTRY table, since it is the primary key in the COUNTRY table.

The data from the soil analytical profiles contain considerably more data than used in the example. All ordered lists of data (tuples) were transferred to dictionary tables to allow establishing referential integrity with fields containing codes in the PLOT or HORIZON tables. A fully normalized database is not necessarily the structure most suited with respect to query performance. Therefore, a database may be implemented in a denormalized form to improve query performance, for example to cater for the needs of Data Warehouses (Shin & Sanders, 2006).

3.3 Amendments to SPADE/M Data Model

The model of the previous SPADE/M database was developed with a close representation of the layout of the forms used in the spreadsheet files of the Proforma II data. It uses two tables to store the parameters measured or observed: one for data describing the plot and one for the horizons of the soil profile. Each parameter is defined as a field in the table. For categorical parameters a dictionary table is linked to the field to explain the codes.

The SPADE/M data model is quite adequate for a relatively small database with a fixed number of parameters and where any additional data fully complies with the specifications of the existing data. Only few modifications to the database model are needed to include the original measurements in the tables of the previous version of the database. The only structural change is an additional field on the data source, which leads to a new primary key for the plot data. The data model with adaptations to the SPADE/M data model to accommodate the new data is shown in Figure 9.

To store the additional measurements the number of fields in the data tables (27; 65) and the number of dictionary tables (13) does not change significantly from the previous version. An advantage of the arrangement is that the representation of the data in the tables resembles a spreadsheet and can be transferred from the database table to a spreadsheet format without much difficulty. The meaning of the entries stored under a field is indicated by the field name. Each field can be formatted individually and referential integrity with dictionary fields can be defined.

It can be argued that the multiple fields for soil texture (6) and water holding capacity (5) should be stored not as individual fields but as a combination of 2 fields, each composed of a value and an identifier for the value parameter. In this arrangement the data have to be stored in separate tables and using a structure different from the main horizon table. For reasons of simplicity and consistency it would appear preferable to continue storing the data in individual fields. More of consequence to data coherence is the use of a single table for measurement methods. It can be ensured that only those methods are recorded with the data which are defined in the methods dictionary, but not that the method associated with a value is appropriate for the parameter. Better coherence between the value reported and the method used could be achieved by either adding a code to identify the parameter or defining separate methods dictionary tables for parameters.

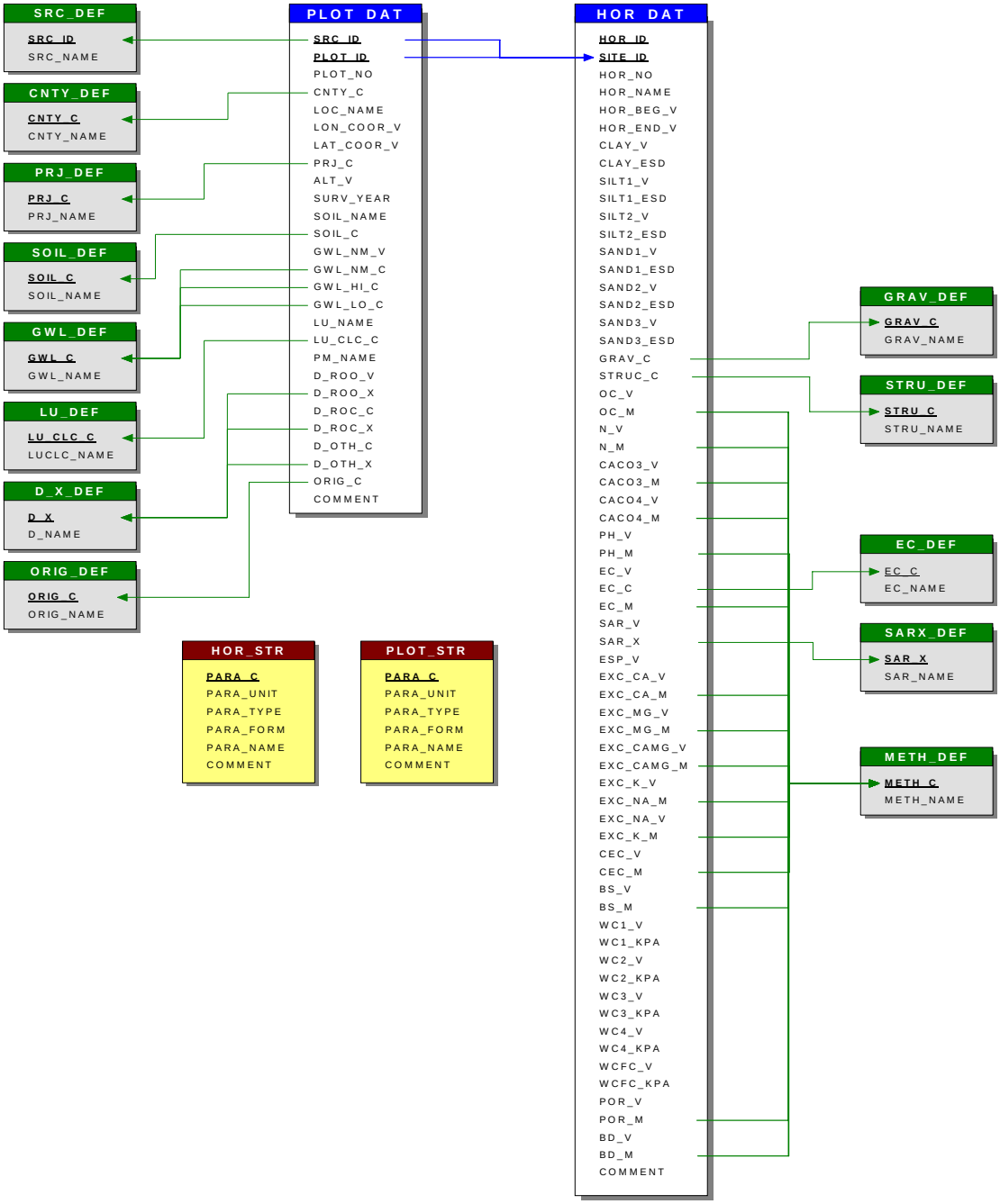


Figure 9: Conceptual SPADE/M Data Model Adapted from Previous Version

The model has other drawbacks when the amount of data is high and when combining data from different surveys and with diverse parameters or measurement units.

- With a considerable amount of data missing from the profiles the arrangement has to deal with empty fields and waste of storage space. Missing data do not *per se* prevent defining referential integrity between tables, but for a more flexible

and compact form of storing the data a complete redesign of the data model is asked for.

- Some of the parameters of the database are not standardized, such as land use or parent material. The parameters are described as text in free form. In the absence of a coding system even minor differences in the description or spelling of the parameter result in problems of finding equivalent conditions when querying the data.
- For high volumes of data, such as generated by the BioSoil survey performed under Forest Focus, any parameter without a value occupies an entry, although the entry may be empty (Null value). These storage requirements of the database are not immediately evident from the file size as stored on the disk, where the file content may be compacted by the DBMS. The actual storage requirements become evident when processing the files. Depending on the DBMS used memory, in RAM or temporary disk space, is set aside to contain the data as dimensioned and this may be beyond the limits of the operating system or the RDBMS. Those requirements can be considerably greater than the file size.
- When integrating data from different surveys with divergent characteristics (parameters, methods, and profile description) the structure of the database has to be adjusted rather than appending the data to the existing tables. The structural changes can lead to adding fields to a table, but may also involve creating additional data and dictionary tables and defining new relations. Adding new fields to a table very likely results in empty entries in the fields of one of the survey data sets and potentially endangers data integrity.

The parameters of SPADE meta-data allow incorporating into a single database a wide range of soil properties using different measurement methods and reporting units. To accommodate data from different surveys a more flexible storage structure was investigated.

3.4 Revised Data Model for SPADE/M2

Given the intervals between surveys of 10 years or more soil databases are generally purposely designed for each survey. The design process is guided by the characteristics of existing data or data to be sampled according the existing specifications. As a consequence, the aspect of having the possibility to frequently adding or modifying data, such as generated by monitoring programs, or integrating various databases has low priority. The design principles for the data model concentrate on avoiding data redundancy (unnecessary duplication) rather than increasing flexibility to accommodate future surveys.

For a more flexible use of the database a complete revision of the model was investigated. The SPADE/M2 data model was developed by identifying the items described by the data, recognizing the relationships between the items and then

associating the items with the data which characterize them. This process is followed by specifying unique identifiers and normalizing the tables.

3.4.1 Items Described by Data

For the revised data model of SPADE/M2 the items or entities described by the data correspond to those of the amended version and are:

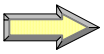
- sampling site;
- soil profile;
- profile horizons.

The new data added to the database use the same items as the data stored in SPADE/M. A set of horizons defines a profile, which is a feature to a sampling site or plot. However, for some plots there are now two descriptions of site conditions, profile and horizons based on the same survey. Therefore, the list of defining attributes for the items consists of the following elements:

- data source;
- plot identifier;
- aggregate method;
- depth segment (horizon).

The defining attributes for SPADE/M2 and their conversion to table format are presented in Figure 10.

SOURCE	PLOT	METHOD	LAYER
A	X1	M1	L1
			L2
			L3
A	X2	M2	L1
			L2
A	X3	M9	L1
			L2
			L3
B	X3	M2	L1
			L2
			L3



SOURCE	PLOT	METHOD	SEGMENT
A	X1	M1	H1
A	X1	M1	H2
A	X1	M1	H3
A	X2	M2	H1
A	X2	M2	H2
A	X3	M9	H1
A	X3	M9	H2
A	X3	M9	H3
B	X3	M3	H1
B	X3	M3	H2
B	X3	M3	H3

Figure 10: SPADE/M2 Key Attributes

The aggregation method corresponds largely to the field "*Origin of data*" in the previous version of the database. It is used to identify a soil profile sampled at a plot as reported by the data source. The depth segment is the identifier of the horizon within a profile. The identifier is ordered according to the depth range of horizons within a profile. It is therefore a combination of the horizon code and the depth range. The horizon codes could not be used on their own because some profiles do not have horizon codes and the depth range is recorded in the database as numeric values giving the beginning and the end of the horizon. The numeric values are stored in a field using float format and hence cannot be used as part of a key. As a consequence the depth segment identifier has to be explicitly declared when entering data.

3.4.2 Relationships between Items

The next step in the design is to identify the relationships between the items. For a soil profile dataset this is shown in Figure 11.

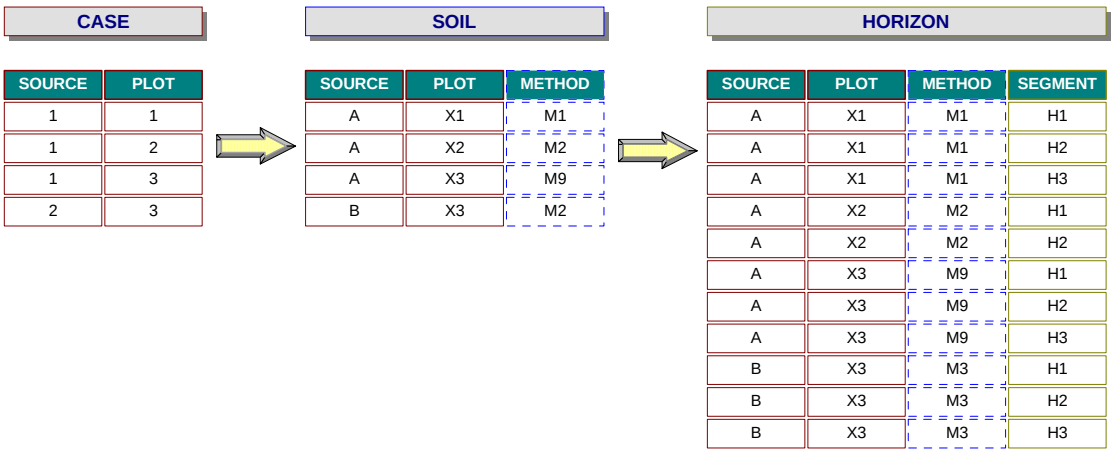


Figure 11: SPADE/M2 Relationship by Main Item

The graph shows the relationship of the attributes between the main items described by the data. With the addition of data for already existing plots and profiles the revision of the data model led to the definition of an extra table on cases of descriptions.

- ***Case***

  In this context the term "*case*" is used to identify the data available for a sample location from a given source. Creating an additional table became necessary because the new data contain not only different data for the horizons of a soil profile but include changes to the data describing the sample location, e.g. land use. The sample location corresponds to the survey plot. The new data refer to existing plots, but with a completely separate description of the plot and associated data. Each case is unique although the referenced sample locations are not.

- ***Soil***

  For a sample location a single set of attributes describing the soil, such as soil type and parent material, is specified. The soil item is a feature of the plot, but unlike other plot observations defined by the horizons. The methods used to describe the soil are the aggregation methods of the soil profile data to define the soil of the plot and the horizon data. Differentiating data by aggregation method is applicable at plot level, but only concerns the soil attribute.

- ***Horizon***

  The horizon table contains the specifications of all horizons. The data are assigned through the soil item to a case. Fore SPADE/M data it is not completely clear whether the data reported for the horizon depend on the aggregation method, i.e. whether the horizon data are the results from more than one profile when such a method is indicated for the soil, or whether data from a

single typical profile are reported. This differentiation of the scope of the aggregation method is not of consequence to the relationship between tables because in SPADE/M only the horizons from one profile instance are recorded. The aggregation method, although possible applicable also to the horizon data, can be included at the level of the soil or plot.

Hence, although there is a soil item there is no separate profile table because each case has just one profile. A separate table would be needed for surveys containing data of more than one profile attributed to a plot from the same source.

3.4.3 Normalization

With the tables and relationships identified between them the task of database normalization can be largely limited to identifying the primary keys for the tables. For the CASE table the primary key is a combination of the case and plot identifiers. For the SOIL table the method field is an element of the candidate key, but the primary key can be formed by a combination of the case and plot identifiers. For SPADE/M2 the aggregation method used to produce the soil data may differ for the same plot, yet is always unique for a combination of plot and source. The primary key for the SOIL table is thus the same as used for the CASE table. As a result, the SOIL table can be removed and any data can be moved to the CASE table. Similarly, the aggregation method is not part of the primary key of the HORIZON table, which uses a combination of case, plot and segment identifiers.

The remaining tables and their primary keys are shown in Figure 12.

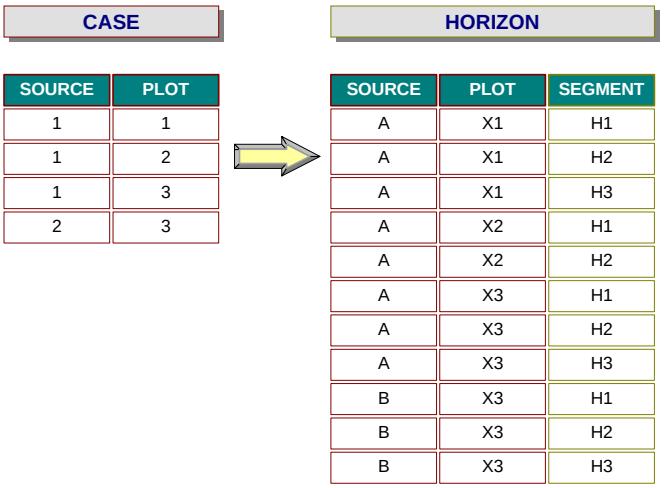


Figure 12: SPADE/M2 Key Attributes by Item, Simplified

The information of the SOIL table and the aggregation method used (field "*Origin*") are treated as attributes to the site location. All horizon data fully depend on the primary key without partial or transitive dependencies.

3.4.4 SPADE/M2 Data Model

Once the relationships between the tables and the primary keys are defined the storage of the fields or attributes can be specified. The model can be described by the following characteristics:

- Data are separated into a database containing the values of measurements and observations and a meta-database explaining the values.
- Tables are linked by identifier fields. In the terminology of the SPADE/M2 database *identifiers* (ID) are members of the primary key and uses to define the relationships between tables. In contrast, *codes* refer to alpha-numeric combinations which convey some meaning of the member of a finite list. Codes are not used in table keys.
- The approach taken to store the plot and horizon data is based on treating all measurements and observations as comprising of a pair of value / parameter, for which only two fields are needed.
- For measured and observed data a distinction is made between categorical and range values and applied to the PLOT and HORIZON tables.
- Any information on data and measurement methods are defined in dictionary tables and possible entries are stored in additional tables. Both elements are part of the meta-database.

The redesigned data model of SPADE/M2 is presented in Figure 13.

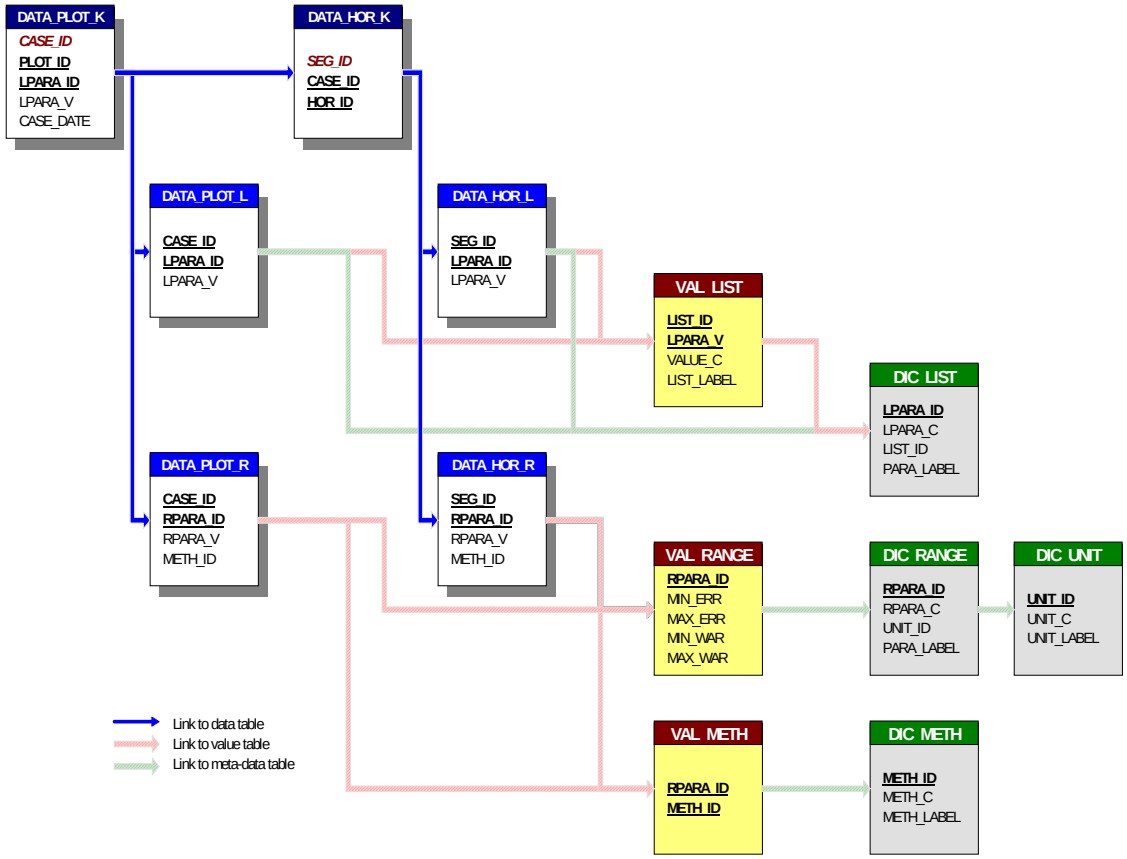


Figure 13: Conceptual Data Model for Soil Profile Analytical Database of Europe of Measured Profiles V. 2 (SPADE/M2)

Details on the database structure are given for the major components.

- ***Data Tables***

  Data tables contain the values from measurements and observations for a plot and profile. For linking the various data tables the revised model introduces tables containing a single field to represent composite primary keys. The governing table for plot records is specified by combining the plot and the source of the data to define the primary key for a case. The source is the value of the parameter with the ID = 1 in the parameter list dictionary. The table contains the parameter ID despite it being constant for all records to maintain an unambiguous link with the parameter dictionary table. The key table for horizon data uses the identifier of the depth segment of a profile to form the primary key together with the case identifier, without repeating the SOURCE and PLOT IDs.

The only reason for this arrangement is a saving in storage space by not repeating the individual fields of the primary keys in the data tables.

The PLOT table contains all data related to the sample location, which includes the soil data and the aggregation method. The horizon data links to the field CADE_ID. An identifier for the depth segment (horizon) is used in combination with the site identifier to form the primary key. The segment identifier should be derived from the horizon code, but as already described neither the horizon code (missing for some sites) nor the depth range (depth stored in field of float format) could be used in the key. Cases of "missing data" represented by a Null value are not included in the data and any instances of missing data need to be specifically coded. Empty entries are present in the field recording measurement methods (METH_ID). For this reason the field cannot be part of a primary key and the soil particle size and water holding capacity parameters cannot be distinguished by the associated method. As a consequence the various individual particle sizes for silt and sand have to be defined as individual parameters rather than a single parameter assessed under different conditions.

- ***Meta Data***

  All explaining information of parameters, methods, units and ranges are stored in tables of meta-data. For categorical data lists of permissible values are given in the corresponding values tables. Applying the reasoning of defining a table of permissible entries also to parameters with real or ratio values has led to the definition of a table of possible ranges of measurements as part of defining the data domain. This information is intended to be used as part of verifying the values and documenting the ranges used.

  For the revised data model the sieve size for the texture data and the suction values when measuring water holding capacity where treated as methods. The corresponding dictionary table was modified by assigning an identifier to each sieve size and suction value for reference. A further consequence is that the various dictionary tables can be merged into a single table.

- ***Separation of Range and Categorical Data***

  In a deviation from other database models the redesign separates plot and horizon data into tables containing categorical and those containing range data. All categorical data are elements of ordered finite lists (tuples) as given in the VAL_LIST table. The corresponding data are codes in either integer or alphanumeric data format. All measurements or observations expressed in measurement units with an infinite number of values on a quantitative scale are treated as range data. Range data are often rational numbers or real numbers, such as $\sqrt{2}$, e or π, but can also be by integers, such as days. Due to the fundamental differences between categorical and range data they are stored in fields of different types. Categorical data are referenced by an ID of type integer, not directly by the codes. Range data are stored in a field of type float.

In the data model the categorical list data are not associated with a measurement methodology. This condition is a consequence of the available data, not an intrinsic property of the parameters. The SPADE/M data contains for categorical plot data the aggregation method of the data for the profile as a whole, but not for the plot or individual parameters. In the absence of a specific profile table the information can only be attached to the plot data as a parameter. The table containing the parameters list links to the table containing the list values by an intermediate field of LIST_IDs. The intermediate field was defined to allow employing one list of values for parameters using the same categories. This is the case for the coding of the groundwater table or the profile depth data.

No measurements units are associated with categorical data because they are generally the result of applying a classification scheme. For range data the measurement unit is associated with the parameter measured, not directly with the measured value. This practice is correct where a parameter is expressed by one and only one unit. When data from different sources are integrated a parameter value can be recorded in another measurement unit. To avoid mixing data recorded in different units it is advisable to store the data as a distinct parameter rather than using one parameter with different units. This approach reduces the potential of mixing data for a parameter expressed in different units.

The concept of a table on possible categorical values (VAL_LIST) has been extended to provide the lower and upper limit of range values. The range limits could have been included in the dictionary table of range parameters as fields. The separate table was created to underline the need for validating the data from those parameters.

Range and list data could be merged into a table of measurements for the plot and one for the horizons. This approach is used for example for the re-designed Forest Focus / ICP Forest Soil Condition database on Level I sites. The data of the key are stored using an alpha-numeric field type. This approach has the advantage of a lower number of tables and less complexity in the relations between tables. When a database does not define IDs for all categorical data but uses alpha-numeric values the ensemble of the measurements and observations can only be recorded in a field of type character. Using alpha-numeric codes to define relationships between tables can lead to inconsistencies depending on the settings on case sensitivity. Importing data from a case-sensitive database into one which is not case-sensitive can therefore lead to key violations. One approach to forestall problems of case sensitivity is to use only upper-case letters for codes. This is, however, not possible when the case is significant to distinguish between situations, as used for the FAO soil type. For example, the FAO74 scheme defines the class *Ch* for *Haplic Chernozem* and FAO90 uses the code *CH* for *Chernozem*. In a non-case sensitive environment the code is treated as duplication and the field cannot be used as a primary key. The SPADE/M2 database uses only upper-case characters in FAO codes and indicates a lower case by preceding the character(s) by an underscore character.

To avoid such problems all fields defining keys and relating tables in SPADE/M2 are based on identifiers (IDs in integer format fields). Therefore, SPADE/M2 measurements and observations could be stored in a field of type float. However, as mentioned before, a field of type float cannot be used to form a key and thus a link to dictionary tables.

The revised data model allows avoiding empty entries and thus avoids using Null values in the measurement tables. Despite the increased inflexibility of integrating soil profile databases with SPADE/M2 there are some intrinsic limitations to consider:

- Defining discrete tables the represent composite primary keys makes for a cluttered model. Distinct key tables are not strictly necessary to relate the data tables and create a field, which is an unaltered combination of existing fields.
- The use of one field containing all measurements and observations restricts linking dictionary tables to a condition. A meaningful link cannot be established by defining relations between tables but depends on procedures to manage the tables and the integrity of the database. Such procedures are not part of the SPADE/M2 database but need to be considered when modifying the profile data or integrating profile data form other sources.
- Separating categorical data from range parameters runs against the concept of looking upon the measurements and observations as attributes describing an entity. It can be argued that the separation follows the nature of the attributes which differ fundamentally and not only by the field type used to store the data. For example, a measurement unit is assigned to range data but not applicable to categorical data while range data are not directly associated with codes unless processed according to a classification scheme. Nevertheless, they could be seen as a conceptual anomaly.
- In SAPDE/M2 a plot identifier is unambiguously related to a sample site on the ground, i.e. the same plot identifier signifies the same sample location whereas different plot IDs signify that data were sampled at different locations. In SPADE/M2 the former profile data could be linked to an existing plot by the identifier noted on the printout. The coordinates indicated on the printouts differed at times from the coordinates given for the plot in the SGDBE. As a consequence, when integrating soil profile data from another source it may not always be evident whether data pertain to a location for which data are also present in another database or to a geographically separate location. As a consequence plots may have been assigned new plot identifiers, but pertain to a location for which data are already recorded in the database. When it cannot be ascertained that two sets of data relate to the same sample location it would be prudent to assign a new identifier to the plot. In the analysis it may be preferable to treat data from the same plot as separate rather then treating data from different locations as coming from the same plot.
- The previous arrangement of declaring parameters as fields is more readily useable in a spreadsheet application. To achieve a comparable arrangement the

tables have to be processed by pivoting the data by the parameter or using a cross-tabulation, which requires additional processing.

- Associating a measurement method at the level of the profile horizon is a result of faithfully translating the Proforma II data. It would be unusual to use more than one method for measuring parameters for the profile horizons. On the forms for measured profiles a single method is at times indicated for the first horizon and not repeated for other horizon data of the profile. One can assume that the method indicated for the first horizon is used for all horizons of the profile. The method should then be an attribute of the profile and should be recorded in the plot table. However, on the Proforma II tables cases occur where the methods used differ between horizons of the same profile.

The database model of SPADE/M2 avoids some problems associated with missing data, is more flexible with respect to accommodating new parameters and provides a more compact storage structure than the previous design. Yet, it constitutes a complete deviation in design and an obstacle to computing values spreadsheet-like along rows.

4 MODEL AND DATA EVALUATION

The evaluation of the SPADE/M2 database model and the soil profile data and aims at assessing the suitability of the database model to store and retrieve data without ambiguity and to provide a measure of the credibility of the parameter values.

4.1 Database Integrity

The database model is evaluated based on an assessment of database integrity. Achieving data integrity is a core requirement of a database design. The aim of data integrity is to avoid introducing inconsistencies into the database when data are inserted, edited or removed from the tables. Three types of data integrity are distinguished[4] depending on the structural element they are concerned with:

- Entity or table integrity: primary keys
- Referential integrity: foreign keys
- Domain integrity: possible entries

Entity and referential database integrity can be enforced by the DBMS used when defining tables and relationships. Specific steps to be taken depend on the system used which maintains data consistency for data storage and retrieval. In SPADE/M2 domain integrity is not enforced by the database model. To achieve domain integrity additional procedures defining the constraints may have to be defined.

When implemented the constraints result in preventing data from entering the database when conditions defined for data integrity are not met. In this case the conditions for data integrity may be modified or the data. Modifying conditions could e.g. be to allow empty entries in a field or to set a default value. To evaluate the database model checks on database integrity were carried out.

[4] http://www.databasedev.co.uk/entity_integrity.html
http://www.databasedev.co.uk/entity_integrity.html
http://www.databasedev.co.uk/entity_integrity.html

4.1.1 Table Integrity

Table or entity integrity refers to the definition of a primary key for each table. The primary keys are shown in the physical database model of SPADE/M2, presented in Figure 14.

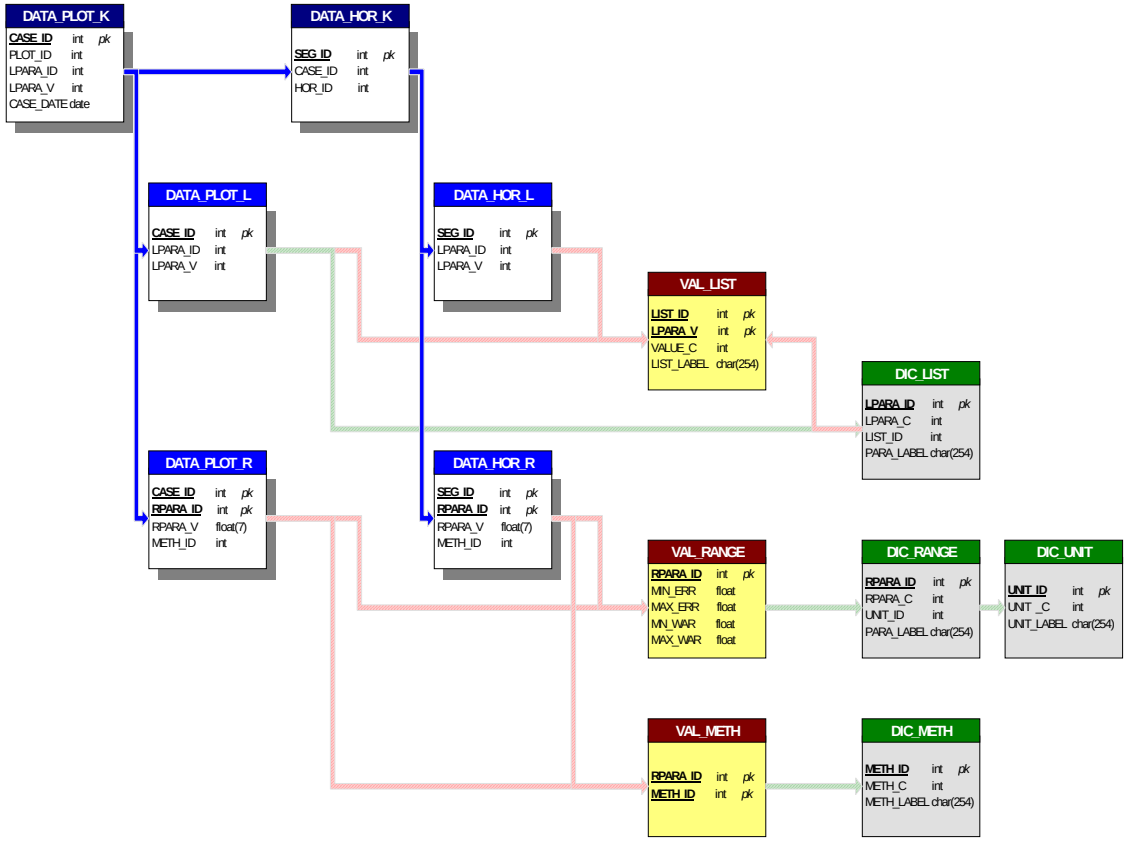


Figure 14: SPADE/M2 Physical Database Model and Meta-Data

Defining primary keys is part of the data normalization process and any database in 2NF should have one primary key defined for each table. Specific demands for the fields of a primary key are enforced by the DBMS when defining the key:

- no empty entries are present in the field(s) forming the primary key (Null);
- the primary key for a record is unique.

These conditions were checked for SPADE/M2 by setting the keys in the development database and listing key violations.

4.1.2 Referential Integrity

Referential integrity sets the rules of relating tables. It means that data of the field one table ("*child*" table) must refer to an existing value in another table ("*parent*" table). The following rules apply to setting referential integrity:

- The field referred to in the parent table must be declared as a *primary key*. This renders the referring field of the child table to contain the values of a primary key not defined in that table, which is therefore a *foreign key* to the child table.
- Referential integrity can only be established between fields of the same type and dimension. Usually, integer and character fields can be used, but not fields of type float.
- Entries in the fields must be identical and care should be taken when setting case sensitivity for character fields.

The treatment of empty entries in the referring field can be set to either allow empty entries or not. For SPADE/M2 the treatment of empty entries depends on the parameter. For example, empty fields are not allowed for the COUNTRY field, because this would impede checking the location of the plots within a country borders. Empty entries are allowed for the field specifying the measurement methodology, which may be unknown.

When referential integrity is introduced to an existing database any entries not conforming to the constraints are usually placed into a specific external table for further evaluation. In case data are entered into a database where referential integrity between tables has been set the new entries are evaluated before entering the database with a message generated by the DBMS. The adherence of a database to the rules for setting referential integrity can also be evaluated by querying the database for the fields concerned.

- ***Coherence between Key and Data Tables***

  The tests for coherence between key and data tables identify plots without a reference in the key table and horizons without a link to an entry in the case table. The completeness of the link fields is ascertained by defining the fields as a component of the primary table keys (no empty entries allowed). The results of the 5 relationships to be tested in the database are given in Table 6.

Table 6: Coherence of Key Tables with Data Values

Table		Correspondence		Status
Parent	*Child*	*Entries*	*Defined*	
DATA_CASE_KEY	DATA_CASE_RANGE	560	560	OK
	DATA_CASE_LIST	560	560	OK
DATA_CASE_KEY	DATA_HOR_KEY	2656	2656	OK
DATA_HOR_KEY	DATA_HOR_RANGE	2656	2656	OK
	DATA_HOR_RANGE	2656	2656	OK

No instances of incoherence between the key and the data tables were found in the database. All cases and horizons are linked to data and all data are structured by the key fields.

- ***Coherence between Data Tables and Values and Dictionary Tables***

  Existence of the categorical plot and horizon values is checked against the entries of the dictionary table listing the parameters and possible values for a parameter given in the values table and the parameters of the dictionary table. When linking the fields of the data table with only the values (LPARA_V) or dictionary table (LPARA_ID) coherence between the entries is only partially evaluated. For a full evaluation all three tables need to be linked.

  For range data coherence concerns the parameters, the methods used and the reporting units. The arrangement of the tables allows checking coherence of entries by linking pairs of tables. The results of the evaluation are given in Table 7.

Table 7: Coherence of Data Tables with Values and Dictionary Tables

Table		Correspondence		Status
Parent	*Child*	*Entries*	*Defined*	
VAL_LIST + DIC_LIST	DATA_PLOT_LIST	5538	5538	OK
	DATA_HOR_LIST	7032	7032	OK
VAL_RANGE	DATA_PLOT_RANGE	2247	2247	OK
	DATA_HOR_RANGE	41598	41598	OK
VAL_METH	DATA_PLOT_RANGE	2247	2247	OK
	DATA_HOR_RANGE	41204	41598	OK
DIC_METH	VAL_METH	158	158	OK
DIC_RANGE	VAL_RANGE	38	38	OK
DIC_UNIT	DIC_RANGE	38	38	OK

For all non-null entries in a child table a correspondence was found in the table referred to.

4.1.3 Domain Integrity

Domain integrity refers the constraints applied to a field with respect to possible or permissible field values. The aspects concerned are:

- field format (type and dimension);
- treatment of empty fields (null support);
- default values;
- special restrictions.

To achieve domain integrity a domain should be declared for all fields. DBMS allow storing settings on field type, null support and default values with the database. Where special restrictions apply additional procedures may have to be used.

- ***Field Format***

  When inserting or editing data directly in a table the DBMS takes care of verifying that only entries conforming to the field type are stored in the table. This inherent check also applies when populating a new structure by importing data from another database. Although the system would not allow that data of an incompatible field type enter the database it is generally advisable to check the conformity of the data to field type specifications by the procedure rather than by the system when committing changes to a table. For SPADE/M2 a procedure

of pre-checking the data imported from SPADE/M was used to evaluate potential incompatibilities between field format limitations and the data.

SPADE/M2 tables use 4 different field types: integer, float(7), character(254) and date. Parameters with logical data, such as SAR_X (Sodium adsorption ratio < 4) were converted to codes to avoid filling all entries with *False* in case the value could not be set to *True* (all entries must be set in a logical field). In a deviation from SPADE/M any fields containing comments or labels are all set to 254 characters. In the previous version the fields were 80 characters wide. The change was made to improve compatibility with other soil profile data, such as sampled by the BioSoil demonstration project. The date of the previous SPADE/M data was set to 29.03.1999, the date of the finalization of the SPADE database as given in the accompanying documentation. The date for the new data was set to 31.12.1995 as during 1995 the profile data for EU-10 countries were completed.[5]

Verified independently of the DBMS used can be field dimensions and data ranges. Those tests should increase the portability of the data between DBMS. The tests for field type conformity applied to the new profile data are given Table 8.

Table 8: Status Options from Test on Field Type

Field Type	Test	Status
Integer[1]	Whole number field with range of 0 to 32,767.	Warning
Number or Float	Numeric entries between 10^{-7} to 10^{7}.	Warning
Character	Characters and/or numbers <255 characters.	Error
Date[2]	Represents a valid date.	Error

[1] For IDs a *long inter* type may have to be used when enlarging the database.
[2] Time fields are not considered in the format verification.

Valid entries for integer fields were checked once for each field. Key IDs are tested in the KEY tables while entries if referring identifiers are checked in the LIST table. This approach should be correct when also checking for data integrity of the database. For the data field the earliest and latest entries are not a qualifying parameter for the check, because all entries have to be valid dates. This verification is done automatically by the DBMS. The results of the check on field formats are summarized in Table 9.

[5] The two dates depend only on the data source (LPARA_V) and are therefore only partially dependent on the combination of fields forming the primary key. The arrangement was maintained to allow subsequent updates of the filed entries, for example to distinguish dates by country.

Table 9: Result for Test on Field Format

Test	Table	Field	Range		Status
			MIN	*MAX*	
IS_INTEGER	DATA_CASE_KEY	CASE_ID	1	560	OK
IS_INTEGER	DATA_CASE_KEY	PLOT_ID	1	560	OK
IS_INTEGER	DATA_HOR_KEY	HOR_ID	1	2656	OK
IS_INTEGER	DATA_HOR_KEY	SEG_ID	1	12	OK
IS_INTEGER	DIC_PARA_LIST	LPARA_ID	1	24	OK
IS_INTEGER	DIC_PARA_RANGE	RPARA_ID	1	44	OK
IS_INTEGER	DIC_METH_ID	METH_ID	0	71	OK
IS_INTEGER	DIC_UNIT_ID	UNIT_ID	0	13	OK
IS_INTEGER	VAL_LIST	LIST_ID	1	24	OK
IS_INTEGER	VAL_RANGE	RPARA_ID	1	44	OK
IS_FLOAT	DATA_CASE_RANGE	RPARA_V	-8.22	3332.0	OK
IS_FLOAT	DATA_HOR_RANGE	RPARA_V	0.0	999.0	OK
IS_CHAR254	DIC_PARA_LIST	PARA_LABEL	4	69	OK
IS_CHAR254	DIC_PARA_RANGE	PARA_LABEL	8	87	OK
IS_CHAR254	DIC_METH	METH_LABEL	8	90	OK
IS_CHAR254	DIC_UNIT	UNIT_LABEL	5	87	OK
IS_DATE	DATA_CASE_KEY	CASE_DATE	NA	NA	OK[1]

[1] Valid date entry checked by DBMS.

Entries in the table VAL_METH are keys from other tables and as such already checked for filed type conformity.

For formatted tables of a database the tests verifying conformity of the values to a data format evaluate the value range rather than a fault of the format. The DBMS takes care that no characters can enter a numeric integer of float field and that date fields only contain valid dates. No further constraints have been defined for the date field, such as limits for a period.

- ***Treatment of Empty Fields***

  All fields included in defining primary keys are set to exclude empty field entries. The settings for other fields differ depending on the connotation of the parameter. In general, fields in dictionary and values tables are set to have non-null entries. Empty field entries are allowed for the following fields:

  - DATA_PLOT_KEY.CASE_DATE
  - DATA_PLOT_RANGE.METH_ID
  - DATA_HOR_RANGE.METH_ID

  There is one exception to the rule for the following field:

- VAL_LIST.LIST_LABEL

Allowing empty entries to the label field of the VAL_LIST table reflects the status of the data in the original files. Forcing the field to exclude empty entries would unnecessarily fill the field with an entry indicating the absent of a value in the source data.

- ***Default Values***

  The database does not set default values for any of the fields.

- ***Special Restriction***

  The database contains some special restriction concerning the structure of the horizons within a soil profile and the threshold for the range values:

  - horizons are numbered sequentially according to the depths of the in the field HOR_START;
  - the depth given in the field HOR_START must be ≤ HOR_END;
  - the depth of HOR_END for horizon (*n*) must be ≤ depth of HOR_START for horizon (*n+1*).
  - VAL_RANGE.MIN_ERR < VAL_RANGE.MAX_ERR
  - VAL_RANGE.MIN_WAR < VAL_RANGE.MAX_WAR
  - VAL_RANGE.MIN_ERR ≤ VAL_RANGE.MIN_WAR
  - VAL_RANGE.MAX_ERR ≤ VAL_RANGE.MAX_WAR

  The rules are not formalized in the database structure and additional procedures have to be used to ensure adherence.

4.2 Data Conformity

Validating the newly entered data and the data imported from SPADE/M into the revised model aims at avoiding introducing erroneous values and inconsistencies into the database. A distinction is made between value inconsistency, i.e. when a value is found outside a predefined range of expected values, and code inconsistency, i.e. when a code is not present in a list. The latter would be checked by the DBMS directly in case conditions for referential integrity were defined.

- ***Value Conformity***

  Value conformity concerns the reliability of range values. For the tests of value conformity a table specifying lower and upper boundaries is included in the database (VAL_RANGE). The table contains the thresholds for errors and

warnings. Any value exceeding an error threshold is considered an impossible or incorrect entry and rejected from the database. The thresholds triggering warnings are set to identify extreme conditions not necessarily outside the range of possible values but with a very low probability.

- ***Code Conformity***

  The tests for codes lead to either accepting the data or an error. Because a code is referred to be an ID in the data table it must exist. The tests therefore cover that the code of the data are associated with a list of suitable entries. The check is closely linked to defining referential integrity between tables. While the structural concept of referential integrity ascertains the existence of corresponding codes the check for code conformity is applied to the codes associated with each parameter and also looks whether the categories defined in the dictionary tables are represented in the data tables.

Where erroneous conditions are found the values are checked against the source data. In case a mistake is found when transferring data between formats, e.g. a typing error, the data are corrected. In case an error cannot be corrected the data are removed from the database. Conditions leading to warnings are checked, but not necessarily rejected.

4.2.1 Single Parameter Tests

The test of valid entries for parameters uses first range values leading to errors and then the set of thresholds testing for possible outliers. The thresholds are the limits of still acceptable values, i.e. are applicable for a comparative test using conditions of less-then (<) and greater-then (>). Using these exclusive operators instead of the inclusive <= or >= operators has the advantage that testing for negative entries is straightforward (x < 0). A disadvantage of the operator is that when a parameter cannot be 0 a minimum value has to be specified as an exclusion criterion.

Table 10: Results on Single Parameter Tests of Value Conformity

Field	Range						Status
	Error		Warning		Data		
	MIN_E	*MAX_E*	*MIN_W*	*MAX_W*	*MIN_D*	*MAX_D*	
ALT	-500.0	8880.0	-150.0	4500.0	-2.00	2600.00	OK
GWL_N_MEAS	0.0	10000.0	0.1	300.0	25.00	750.00	OK
D_ROO	0.0	10000.0	0.1	300.0	5.00	200.00	OK
D_ROC	0.0	10000.0	0.1	300.0	**0.00**	235.00	**Warning**
D_OTHOBS	0.0	10000.0	0.1	300.0	**0.00**	235.00	**Warning**
HOR_START	0.0	10000.0	0.0	300.0	**0.00**	235.00	**Warning**
HOR_END	0.0	10000.0	0.1	1000.0	**0.00**	999.00	**Warning**
CLAY	0.0	100.0	0.0	99.0	0.00	81.00	OK
SILT_1	0.0	100.0	0.0	99.0	0.00	77.00	OK
SILT_2	0.0	100.0	0.0	99.0	0.00	92.00	OK
SAND_1	0.0	100.0	0.0	100.0	0.00	96.00	OK
SAND_2	0.0	100.0	0.0	100.0	0.00	92.50	OK
SAND_3	0.0	100.0	0.0	100.0	0.00	100.0	OK
ORG_C	0.0	80.0	0.0	70.0	0.00	56.80	OK
ORG_M	0.0	100.0	0.0	100.0	0.00	**109.0**	**Error**
N_TOT	0.0	100.0	0.0	25.0	0.00	21.50	OK
CACO3	0.0	100.0	0.0	100.0	0.00	99.00	OK
CASO4	0.0	100.0	0.0	50.0	0.00	36.00	OK
PH	0.0	14.0	2.5	12.5	2.60	10.0	OK
EC	0.0	100.0	0.0	60.0	0.00	53.60	OK
AR_NA	0.0	100.0	0.0	50.0	0.00	49.00	OK
EXCH_NA_P	0.0	100.0	0.0	75.0	0.00	73.00	OK
EXCH_CA	0.0	1000.0	0.0	115.0	0.00	110.80	OK
EXCH_MG	0.0	500.0	0.0	50.0	0.00	47.60	OK
EXCH_CAMG	0.0	1500.0	0.0	150.0	3.00	91.00	OK
EXCH_K	0.0	100.0	0.0	50.0	0.00	5.00	OK
EXCH_NA	0.0	500.0	0.0	100.0	0.00	65.80	OK
CEC	0.0	1000.0	0.0	200.0	0.20	179.60	OK
BS	0.0	100.0	0.0	100.0	0.20	100.00	OK
WC_1	0.0	100.0	0.0	95.0	0.00	**183.00**	**Error**
WC_2	0.0	100.0	0.0	95.0	0.50	**140.00**	**Error**
WC_3	0.0	100.0	0.0	95.0	0.50	**112.00**	**Error**
WC_4	0.0	100.0	0.0	95.0	0.00	94.30	OK
WC_FC	0.0	100.0	0.0	95.0	0.00	92.90	OK
POR_TOT	0.0	100.0	10.0	95.0	19.00	94.00	OK
BD	0.0	3.0	0.05	2.3	0.08	2.15	OK

The conformity check for single parameter ranges revealed some conditions leading to warnings or errors. Particular problems encountered were:

- ***Depth to Obstruction to Rooting***

  For soil profiles data the start of an obstruction to rooting should be below the surface or there would hardly be any soil. For 10 plots located in the Slovak Republic the depth rock obstruction (D_ROC) is given as 0 cm. The depth to an obstruction to rooting other than rock (D_OTH) was given as 0cm for 15 plots. For all plots a value of the depth of rooting (D_ROO) other than 0 was given. For the 5 overlapping cases the same non-zero value was given for D_ROO and D_ROC. It would appear that the value 0 for D_ROC and D_OTH were used to indicate the absence of a measurement rather than the start of an obstruction at the profile surface.

- ***Zero as Minimum Parameter Value***

  A general problem is recording zero (0) entries. In most cases they seem to signify a record of the absence of a measurement instead of the result of a measurement where the parameter was not found. The meaning of a zero entry for a parameter could not be determined with certainty and a minimum threshold of zero was accepted.

- ***HOR_END = 0***

  For one horizon of a profile in Portugal (Plot code: "*Bh-2*") only a single depth value (0) was given for the organic layer. It was assumes that the layer was probably less than 1cm in height and the horizon was retained.

- ***Water Retention Value (WC) > 100***

  In 3 cases a water retention value > 100 was found. The values were treated as erroneous entries, because the unit for the parameter is volume percentage of water and therefore cannot exceed a value of 100.0. All cases were recorded for the first horizon of a profile in Romania (Soil name: "*Unknown3*") in the RO.XLS file. The data were subsequently removed from the table of SPADE/M2.

- ***OM > 100.0***

  For one profile in the UK (Plot: 349) with a *histosol* (*Od*) the values reported for organic matter exceeded 100.0. For this profile no data on the actual measurement results were available. The organic matter values were computed from measurements of organic carbon by using a fixed conversion factor. The conversion factor varies depending on the nature of the organic material and in this case it was assumed that a different value should have been used. The values were reset the 100.0 in the database.

For the lower limit of a horizon a value of 999 is used to code the situation that the limit was not determined. The threshold for giving a warning could have been set below the

value to indicate the condition that the value does not represent a measured depth but a code for a condition.

4.3 Multiple-Parameter Tests

The conformity of some values depends on the conditions defined by one or more other parameters. In SPADE/M2 multi-parameter tests are not part of the database design and have to be implemented as procedures. This makes inserting data into the database more complex. While single-parameter tests can be implemented at the time of data entry and erroneous values do not enter the database a status of a value subjected to a multi-parameter test can only be reliably determined when all relevant data have been entered into the database. Until the tests have been applied the database remains in an intermediate state because data may subsequently be modified or removed. For DBMS without an intermediate storage of data before committing any changes using two databases, one for working and one for publishing, can be a workable option.

Table 11: Multi-Parameter Checks

Condition	Range						Status
	Error		Warning		Data		
	MIN_E	*MAX_E*	*MIN_W*	*MAX_W*	*MIN_D*	*MAX_D*	
COOR_X where PROJ = GEO	-180.0	180.0	-11.0	45.0	-8.22	29.58	OK
COOR_Y where PROJ = GEO	-90.0	90.0	34.0	72.0	4.28	59.47	**Warning**
D_ROC – D_ROO	300.0	10000.0	0.0	300.0	-200.00	150.00	**Warning**
D_OTH – D_ROO	300.0	10000.0	0.0	300.0	-200.00	30.00	**Warning**
CASE_ID where MIN(HOR_START)	0.0	-	-	-	0.00	-	OK
CASE_ID where 0 - MIN(HOR_START)	-300.0	-	0.0	-	-10.00	-	**Warning**
HOR_END – HOR_START	300.0	10000.0	0.0	300.0	0.00	190.00	**Warning**
HOR_END(n) – HOR_START(n+1)	300.0	10000.0	0.0	0.1	90.00	90.00	**Warning**
CLAY + SILT(n) + SAND(n)	0.0	110.0	0.0	100.0	0.00	120.25	**Error**

Note: depth value 999 excluded from computations.

Problems found in the data from testing multiple parameters were:

- ***Geographic Coverage: COOR_Y***

  Coordinate limits for warnings are set to cover mainland Europe plus Turkey when using geographic coordinates. Not included are overseas areas and islands. The profile with the offending y-coordinate is located on the Canary Islands and the warning can be ignored.

- ***Rooting Depth and Depth to Obstacle by Rock: D_ROC – D_ROO***

  The difference between the depth of the rock obstacle to rooting and the rooting depth should be >= 0. For 11 profiles a negative difference was found, 10 located in the Slovak Republic and for 1 profile in the UK (Case ID: 513, Plot 452). For the profiles of the Slovak Republic all occurrences concern cases where D_ROC was set to 0. The data for the UK profile were reported as found in the database. The previous data for the plot did not contain information of rooting depth or depth to rooting from rock.

- ***Rooting Depth and Depth to Obstacle other than Rock: D_OTH – D_ROO***

  The difference between the depth to rooting by an obstacle other than rock and the rooting depth is negative for 20 profiles. In 5 cases a value other than 0 but less than D_ROO was given for D_OTH. For 2 newly entered profiles (CASE_ID: 531, 553) the entries were verified with the hard-copy data and found correctly transferred. For 3 profiles located in Romania (CASE_ID: 271, 272, 292) the values given for D_OTH were also > 0. For the remaining profiles with a negative difference (CASE_ID: 205, 206, 207, 208, 209, 210, 211, 212, 218, 215, 216, 217, 218, 219, 220) the negative difference was caused by entries of 0 in the D_OTH field.

- ***Horizon Depth Consistency: HOR_END – HOR_START***

  There were no negative differences between the start of a horizon and the end but for 6 horizons the start an end values were identical, all from different profiles (CASE_ID/SEG_ID: 77/2, 259/1, 343/6, 344/5, 345/3, 553/6). Where it was possible (CASE_ID: 553) the data were verified on the hard-copy and found to be correctly transferred tot the database. No obvious reason for using identical values for horizon start and end could be identified.

  Also tested with the horizon depth was the start of the first horizon. For the SPADE/M data no negative values for horizon depth are allowed, which includes organic layers over mineral layers. Any horizons with a negative depth value would therefore result in an error condition. None were present in the database. Another condition to test is that a profile is describes fro a depth of 0cm downwards. Any profile with the start of the topmost horizon starting at a lower depth ([0 – MIN(*HOR_START*)] < 0) leads to a warning status. This condition was found for 3 profiles (CASE_ID: 56, 173, 325). A reason for the absence of a horizon starting at the surface could not be established.

- ***Horizon Depth Continuity: HOR_END(n) – HOR_START(n+1)***

  The continuity of horizons in a profile is tested for gaps and overlaps. Gaps occur when part of a profile is not defined as a horizon (HOR_END(*n*) – HOR_START(*n+1*) < 0). Horizon overlaps occur when part of the profile is described by more than one horizon (HOR_END(*n*) – HOR_START(*n+1*) > 0).

  Horizon gaps were found for 13 plots (CASE_ID: 54, 56, 325, 384, 413, 414, 415, 419, 421, 426, 427, 432, 435). The occurrences were checked for transcription errors and found to be represented in the database as found in the source data.

  Overlaps were signaled for 17 plots (CASE_ID: 376, 401, 406, 407, 413, 414, 415, 416, 417, 419, 420, 421, 424, 426, 427, 432, 435). The reason for overlaps of horizons in the position within a profile is the use of different depth values for the first and second table parts in the spreadsheet files. The two parts should represent the same horizons which is specified in the first part, although the depth parameter is repeated in the second part.

  Most of the differences highlighted by the test relate to sites in Switzerland. For most of the plots only a single value for depth is given instead of a range or a single value in the first part of the table and a range in the second. Even using user analysis in the instances reported it was not possible to define coherent horizon limits between the two parts of the table. There is not obvious reason for reporting different depths between the parts.

- ***Sum of Particle Size (CLAY_CONT + SILT(n)_CONT + SAND(n)_CONT)***

  The sum of the relative content all particle sizes should be 100.0%. Rounding or inaccuracies in measuring can lead to very minor deviations from that value. For *histosols* the content for sand or clay particles is at times stated and the sum of the particle sizes would then be considerably less than 100.0%. In the database a sum of all particle sizes of, 100.0% was found for 409 horizons, while a sum> 100.0% was computed for 354 horizons. In some cases of sums exceed 100.0% by a considerable margin quite outside the expected range of measurement uncertainties. The test uses a strict value of 100.0 as a threshold. Depending on the application of the data a more lenient threshold to account for measurement uncertainties and rounding could be used.

4.4 Code Conformity

For an evaluation of the conformity of codes the occurrence of a listed value with respect to the range of codes defined in the value list and dictionary tables is assessed for each parameter.

4.4.1 Parameter Data in Value List

The range of possible entries of categorical values in the data table is defined by the VAL_LIST table. The existence of the list items used in the data tables in the list of possible items has already been established when evaluating referential integrity. The check assesses how many of the possible list items are actually present in the data tables for each parameter. The results of the tests applied are given in Table 12.

Table 12: Coherence of Data List with Values List (VAL_LIST)

List Item		Correspondence		Status
Table	*Parameter*	*Entries*	*Defined*	
DATA_CASE_KEY	SOURCE	2	2	OK
DATA_CASE_LIST	COUNTRY	18	18	OK
	LOC_NAME	184	560	OK
	PROJ	4	7	OK
	SOIL	560	560	OK
	FAO	139	448	OK[1]
	GWL_NM	5	5	OK
	GWL_HI	5	5	OK
	GWL_LO	5	5	OK
	LU	484	560	OK
	PM	492	560	OK
	D_ROO_X	2	2	OK
	D_ROC_X	2	2	OK
	ORIGIN	3	6	OK
	COMM_PLOT	195	560	OK
DATA_HOR_LIST	HOR_NAME	482	482	OK
	GRAVEL	6	6	OK
	STRUCT	10	10	OK
	AR_NA_X	1	1	OK

[1] Code *LVc* was not found in the FAO90 meta-data. A corresponding entry was added to table VAL_LIST.

In the table the number of *defined* correspondences refers to number of entries which are defined in the list of values tables, but not necessarily indicating a meaningful entry.

- ***Coding Potential Occurrence***

  For several parameters individual codes are defined for each plot while the explaining field is empty, acting as a placeholder. This situation occurs for the parameters for the soil name given by the source (SOIL), the indication of land use (LU) and the parent material (PM). For those parameters the entries are not coded according to a classification scheme, although the name may indicate this, resulting in a 1:1 relationship between data and list entries. In case additional information on the parameter value becomes available a record has already been created to receive the data

- ***Coding Reported Occurrence***

  A different approach has been taken for coding names of horizons (HOR_NAME). The values list contains the codes for names for all horizons, for which a name was recorded. The names were coded according to exact matches and using case-sensitivity. As a consequence, a 1:n relationship exists between the fields [VAL_LIST.HOR_NAME] and [DATA_HOR_LIST.HOR_NAME]. For new horizon data a new record has to be created in case the new code does not match an existing entry.

- ***FAO Soil Name***

  For FAO soil codes a substantial difference between the codes defined by the classification scheme (448) and the occurrence of the codes in the data tables (139) is noted. This situation does not signify that codes from only FAO74 or FAO90 were used for the profiles. FAO74 codes were used to describe the soil type for 442 plots, while FAO90 codes were used for 118 plots.

4.4.2 Conformity of Measurement Methods

For range parameters the method of how a parameter was measured is associated with each value. All methods are defined in a single comparable to the value list for category parameters parameter. The results of testing conformity of the method associated with a measured value are presented in Table 13.

Table 13: Conformity of Parameter Methods with Possible Methods

Range Item		Instances	Status
Field	*Method*	*No.*	
COOR_X	Unknown assessment method[1]	472	OK
COOR_Y	Unknown assessment method[1]	472	OK
ALT	Unknown assessment method[1]	486	OK
GWL_N_MEAS	Unknown assessment method[1]	237	OK
D_ROO	Unknown assessment method[1]	288	OK
D_ROC	Unknown assessment method[1]	171	OK
D_OTHOBS	Unknown assessment method[1]	121	OK
SILT_2	Particle size measured at 2 micro meter	1	**Error**
SAND_1	Particle size measured at 50 micro meter	1	**Error**
SAND_1	Particle size measured at 75 micro meter	6	**Error**
SAND_1	Particle size measured at 90 micro meter	5	**Error**
SAND_1	Particle size measured at 100 micro meter	44	**Error**
SAND_1	Particle size measured at 105 micro meter	89	**Error**
SAND_1	Particle size measured at 125 micro meter	5	**Error**
SAND_1	Particle size measured at 150 micro meter	15	**Error**
SAND_2	Particle size measured at 105 micro meter	6	**Error**
SAND_2	Particle size measured at 150 micro meter	17	**Error**
ORG_C	Wet digestion (Kjeldahl method) (%)	109	**Error**
ORG_C	Other method for total nitrogen	5	**Error**
ORG_M	Method of Walkley and Black	71	**Error**
N_TOT	Method of Walkley and Black	3	**Error**
N_TOT	Other method for organic carbon, to be specified separately	2	**Error**
N_TOT	Calcimeter method (%), measures CO2 emitted	1	**Error**
CEC	Neutral Ammonium Acetate (NH4AOc) extract, cmol+/kg	3	**Error**
BD	Other method for total pore space	11	**Error**

[1] No method indicated in source data.

For parameter associated with the plot no methods were recorded in the Proforma II files. The corresponding code (A00) was associated with those data. The test resulted in a status of "OK" by design.

For measurements methods of range parameters only those conditions are shown, for which problems of conformity between measurement and method were found. Specific problems in associating methods to parameter values were:

- ***Particle Size***

  For one horizon the particle size of the sieve for determining silt was given as 2µm (CASE_ID: 192). This size is correctly used for the clay component of the profile and a value of 200 µm is given for the sand component. The particle

components sum up to 100.0%, which makes it unlikely that the value for the silt portion is incorrect and the value could be retained in calculations.

For the sand component an inappropriate method for determining the particle is given for 188 horizons. According to the description of the SPADE database the size of the sieve to determine the sand component should be $\geq 200\mu m$. Concerned are horizons of 88 plots, usually all horizons of the plots. This indicates that the sizes given are not inadvertently entered but more likely correspond to the sizes used to determine the sand portion.

- ***Organic Carbon***

  For 114 horizons of 34 plots a method for measuring nitrogen instead of organic carbon was recorded. In all cases the method for measuring nitrogen was "*Other method for total nitrogen*" (A05). Except for two plots in Spain the situation occurs on practically all French plots. The values for OC are within the expected ranges for the horizons and no indications for not including the data in further computations were found.

- ***Organic Matter***

  For organic matter the method "Method of Walkley and Black" is indicated for 26 plots in the UK. The method is appropriate for determining organic carbon in the soil. The method would be correct for determining OC but was highlighted for OM because the factor used to convert organic carbon into organic matter is not specified by a method.

- ***Total Nitrogen***

  For 5 horizons the method given for total nitrogen is used to determine organic carbon and for 1 horizon for determining $CaCO_3$. A possible explanation for the incoherence is that the parameters were recorded in neighbouring fields in the Proforma II table, similar to the method for determining Nitrogen associated with organic carbon (see above).

- ***Cation Exchange Capacity***

  An inappropriate method for determining CEC was given for 3 horizons of a single plot ("*Neutral Ammonium Acetate (NH4AOc) extract, cmol+/kg*" (A27); CASE_ID: 132). The corresponding value (0.20) is the same for all 3 horizons and the lowest value of all profiles in the database.

- ***Bulk Density***

  An incorrect method for determining bulk density was given for 11 horizons belonging to 3 profiles. In all cases a method for "*Other method for total pore space*" (A27) was attributed to the measurements. The values for bulk density are within the accepted range and present no cause for not including the bulk density values in computations.

5 SUMMARY

Soil profile data from 64 profiles located in England and Wales, 12 of which are for new plot locations, could be added to the SPADE/M database. A significant element of the recovered data is the geographic position of the plots, which was so far not available for the profiles. For plot parameters the recovered Proforma II tables contain information which matches the data stored in the spreadsheet tables of the SGDBE and SPADE/M, including the soil type. However, significant differences between the data sources were found for the parameters describing the profile. The differences were attributed to the condition that the SGDBE soil profile data for England and Wales describe a single representative profile while the recovered data are the average of a number of measured profiles. An issue for further discussion is whether representative profile data can be meaningfully linked to a point location.

The recovered data were only available in form of printouts. The corresponding electronic files could not be located. Therefore, all data were entered manually into a temporary structure. To verify that the values were transferred reliably a procedure of validating the entries was developed. This procedure is based on an approach based on a stepwise assessment of the data of first excluding any values outside possible ranges and then evaluating the data according to the probability of occurring. The validation process was extended to the data from the previous SPADE/M database to form a common base.

A data evaluation procedure was applied to a revised database model for the measured profile data. The main aim of the database revision was to facilitate the integration of the SPADE/M profile data with other soil profile database. Data integration should be achieved by modifying tables and without the need of restructuring the database. The revise model stores plot and profile data as a pair consisting of the measured value and the referring parameter. The meta-database has been considerably extended to facilitate checks for data conformity. For categorical parameters the list of possible entries is defined in a specific table. For range values the thresholds used validate the entries has been added to the database. Table settings ensuring data integrity were verified by using query procedures as detailed in this document. Such procedures can be useful in an environment where only the data tables are distributed and all settings for data integrity are not stores with the tables.

The revised database model separates categorical and range data into distinct tables. The separation follows the characteristics of the data and relationship. This treatment runs to some degree contrary to the approach of treating a measurement as an attribute of an entity. The setup of separating the parameters is more complex than a unified approach but provides more control over the data from within the database, thus reducing the need for additional procedures. Yet, in practice it may prove unnecessarily convoluted and for the integration of validated databases and where data are not updated frequently. The model is experimental and the flexibility of integrating other profile data should be tested.

Acknowledgement:

Measured profile data remain valuable over the years to provide indicators of past conditions of our environment and are worth the effort of preserving such information. The author wishes to express his thanks to R.J.A. Jones and J.M. Hollis form National Soil Resources Institute, Cranfield University for their perseverance over many years to recover the tables on measured profiles and safe them from being lost. Further thanks go to J. Daroussin (INRA - UR SOLS, Orleans) for his verification of the coordinate transformation of the French profiles.

References

Bouron, P. (2005) Cartographie – Lecture de Carte. Ecole Nationale des Sciences Geographiques. 6 et 8 avenue Blaise Pascal, 77455 Marne la Vallée Cedex 2. www.ensg.ign.fr. 101pp.

Breuning-Madsen, H. and Jones, R. J. A. (1995): Soil profile analytical database for the European Union. Geografisk Tidsskrift, Danish Journal of Geography (95):49-57.

Breuning-Madsen, H. and R.J.A. Jones (1998): Towards a European Soil Profile Analytical Database. pp. 43-50. in: Heineke, H.J., W. Eckelmann, A.J. Thomasson, R.J.A. Jones, L. Montanarella and B. Buckley (eds.): Land Information Systems: Developments for planning the sustainable use of land resources. European Soil Bureau Research Report No.4. EUR 17729 EN. Office for Official Publications of the European Communities, Luxembourg.

Codd, E.F. (1970) A relational model of data for large shared data banks. Communications of the ACM. 13(6). p. 377-387.

Darwen, H. (2009) Windows Control Panel Regional Settings Properties. Hugh Darwen and Ventus Publishing ApS. Download from www.BookNooN.com. 230pp.

Durrant-Houston, T. and R. Hiederer (2009) Applying quality assurance procedures to environmental monitoring data: a case study. J. Environ. Monitor., 2009, 11, pp774 – 781.

Gilfillan, I. (2002) Introduction to Relational Databases. Database Journal. 24.06.2002. http://www.databasejournal.com/sqletc/article.php/1469521/Introduction-to-Relational-Databases.htm

Hiederer, R., R.J.A. Jones and J. Daroussin (2006) Soil Profile Analytical Database for Europe (SPADE): Reconstruction and Validation of the Measured Data (SPADE/M). Geografisk Tidsskrift, Danish Journal of Geography 106(1). p. 71-85.

Hiederer, R., T. Durrant, O. Granke, M. Lambotte, M. Lorenz, B. Mignon, and V. Mues (2007) Forest Focus Monitoring Database System – Validation Methodology. EUR 23020 EN. Office for Official Publications of the European Communities, Luxembourg. 56pp.

Howard, P. J. A and D. M. Howard (1990) Use of organic carbon and loss-on-ignition to estimate soil organic matter in different soil types and horizons. Biology and Fertility of Soils 9(4). p. 306-310.

Shin, S.K. and G. L. Sanders (2006) Denormalization strategies for data retrieval from data warehouses. Decision Support Systems, 42(1). p.267-282.

Data Update and Model Revision for Soil Profile Analytical Database of Europe of Measured Parameters (SPADE/M2)

European Commission

EUR 24333 EN – Joint Research Centre – Institute for Environment and Sustainability

Title: Data Update and Model Revision for Soil Profile Analytical Database of Europe of Measured Parameters (SPADE/M2)

Author(s): R. Hiederer

Luxembourg: Office for Official Publications of the European Communities

2010 – 55 pp. – 21.0 x 29.7 cm

EUR – Scientific and Technical Research series – ISSN 1018-5593

ISBN 978-92-79-15646-5

DOI 10.2788/85262

Abstract

The Soil Profile Analytical Database of Europe of Measured parameters (SPADE/M) is part of the distribution package of the Soil Geographic Database of Eurasia (SGDBE). Typical combinations of profile parameters and morphological characteristics of the sample site were intended to support the definition of generalized rules for estimating pedological and hydrological properties of the pedo-transfer rule (PTR) database of the SGDBE. In 2005 the data of the SGDBE were transferred to a common data storage structure. In 2008 original hard-copies on profile measurements were re-discovered at the National Soil Resources Institute, Cranfield University (NSRI). To make the original data more generally available the profiles were added to the existing database. This step required changes to the structure of the database and a validation of the all entries for accurate and reliable data storage and retrieval.

Data Update and Model Revision for Soil Profile Analytical Database of Europe of Measured Parameters (SPADE/M2)

How to obtain EU publications

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

LB-NA-24333-EN-C

JRC
EUROPEAN COMMISSION

ISBN 978-92-79-15646-5

Publications Office
Publications.europa.eu